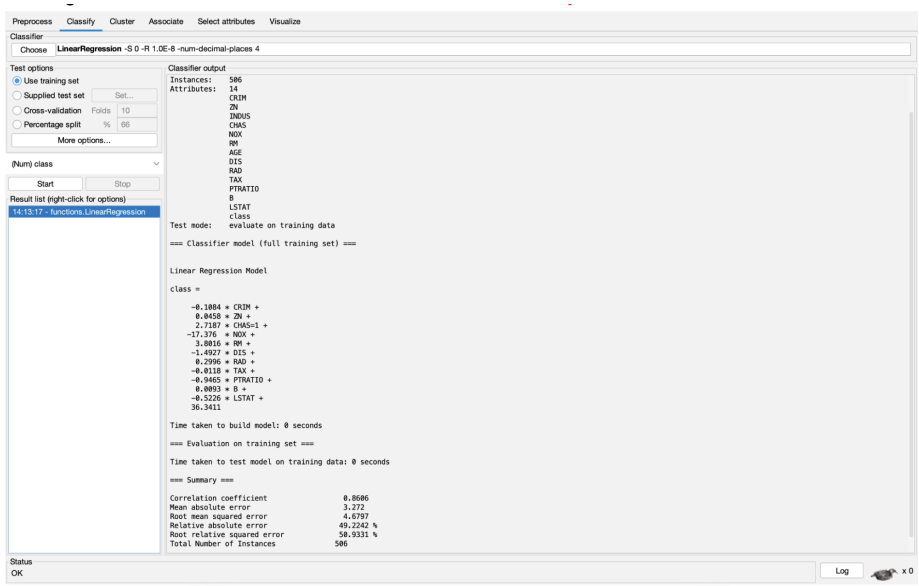
Data Mining Rubrics

WEKA

Q1 (2 points). Build a linear regression equation to predict MEDV. Include a screenshot that shows the linear equation. How many terms are in the equation, and 'why'? In other words, discuss the resulting equation.

POINTS	DESCRIPTION
1 point	For the screenshot: The regression model should be in the screenshot. *If not, 0 points*. 
1 point	For the description of results, each following point is worth 0.5 points: a) Number of terms. By default, if we use all 14 attributes (including the class attributes), the number of terms should be 12 (including the constant) or 11 (except the constant). If the student has deliberately deleted some attribute, it is acceptable if they have explained their reasons in the README. If not, 0 points. b) A reasonable discussion about any one of the following points: i) The “missing” attributes INDUS and AGE; one acceptable reason is that WEKA discards statistically insignificant attributes as measured by their R-squared

	value ii) The weight of each attribute (negative weight means a higher value of the attribute usually lowers the median home price whereas the one with a positive weight raises it) iii) Other rational interpretations of the equation

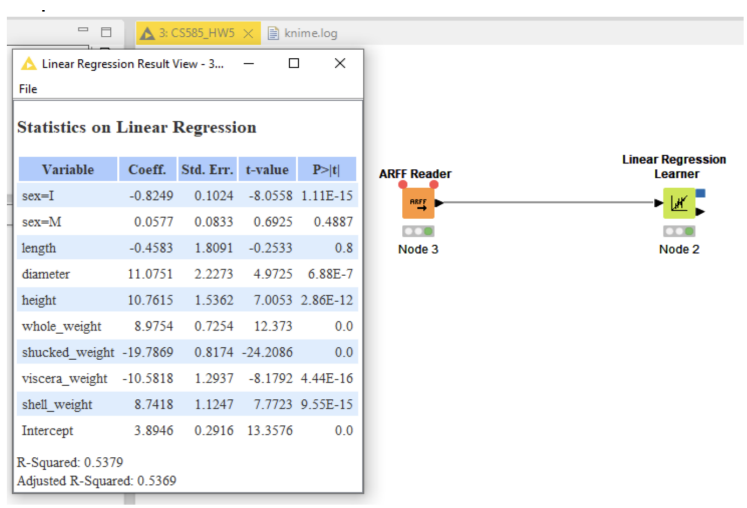
*** if the information on the screenshot does not match the discussion, 0 points for Q1

KNIME

Q2 (2 points). Use KNIME to perform linear regression [on all parameters, not a subset]. You need these nodes: AARF Reader, Linear Regression Learner. Create and connect the nodes, and execute each. What is the linear equation? Include a screenshot.

POINTS	DESCRIPTION
1 point	Screenshot shows workflow (students may use a different file reader but not a different learner)
1 point	Screenshot shows statistics - At least the first decimal of each coefficient must match the values below eg. the coefficient of length may be 0.46 but not 0.5. If the coefficients vary drastically and the student has not explained the modifications they made, 0 points - If any one variable is missing, 0 points

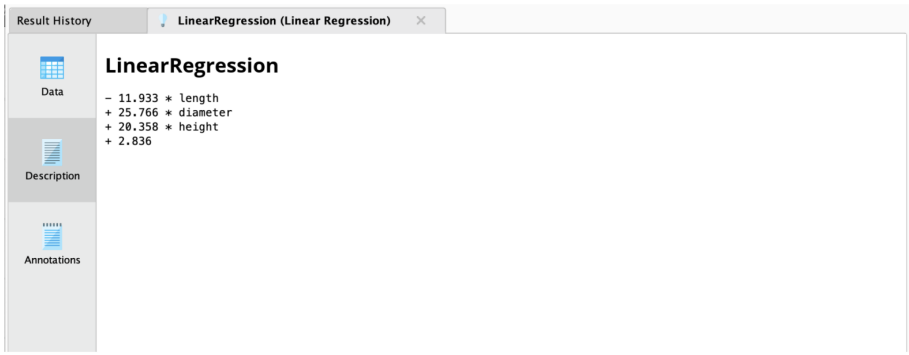
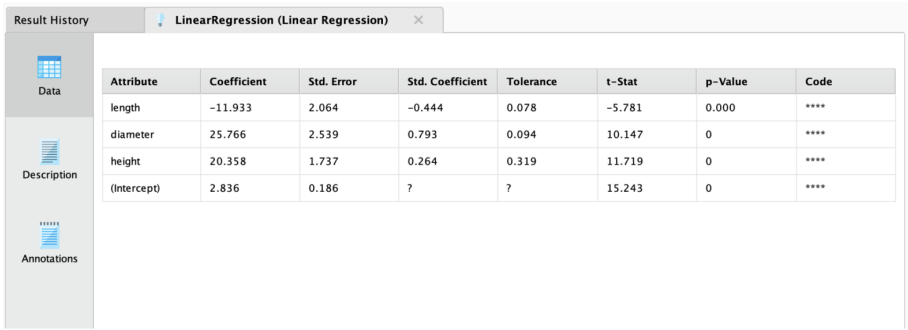
Sample screenshot:

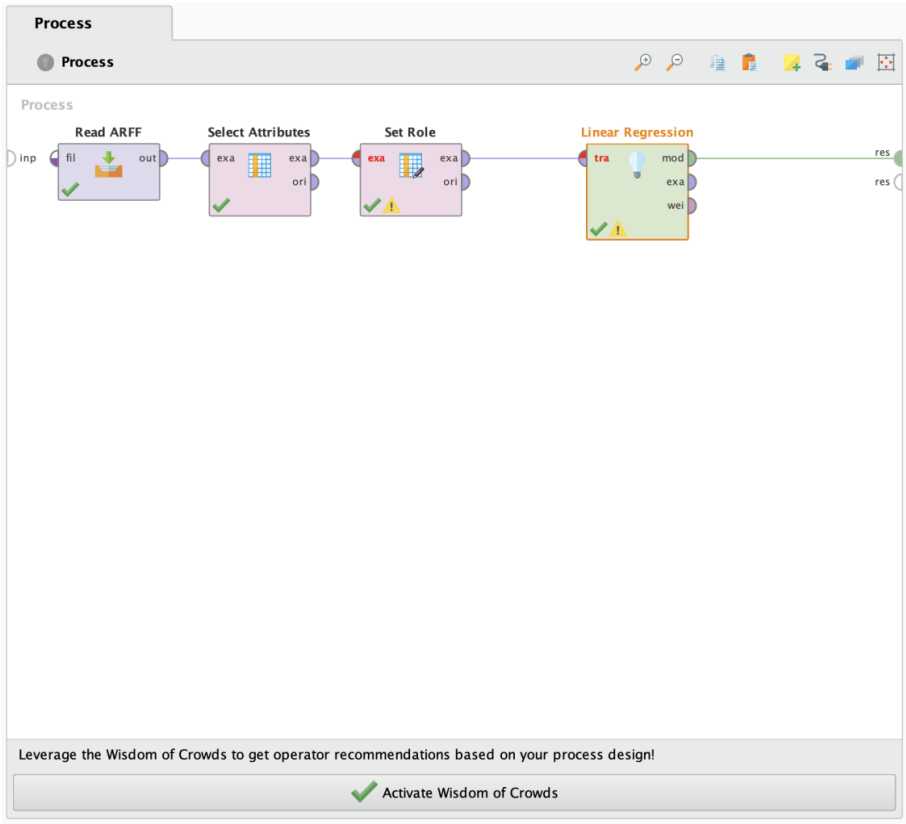


Variable	Coeff.	Std. Err.	t-value	P>|t|
sex=I	-0.8249	0.1024	-8.0558	1.11E-15
sex=M	0.0577	0.0833	0.6925	0.4887
length	-0.4583	1.8091	-0.2533	0.8
diameter	11.0751	2.2273	4.9725	6.88E-7
height	10.7615	1.5362	7.0053	2.86E-12
whole_weight	8.9754	0.7254	12.373	0.0
shucked_weight	-19.7869	0.8174	-24.2086	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

RapidMiner

Q3 (1.5 points). Bring in the shells.arff data, and only work with these 4 params: length, diameter, height,num_rings. Do a linear regression to predict num_rings, from length, diameter, and height. Question: what is the equation? Include a screenshot. Note that you need a 'Set Role' node where you would set num_rings to be a "label", before doing the regression. The regression itself would be done using a 'Linear Regression' operator.

POINTS	DESCRIPTION
1 point	The equation in README.txt: `num_rings = - 11.933 * length + 25.766 * diameter + 20.358 * height + 2.836` OR a screenshot of it: 
0.5 point	Screenshot showing the student's work (any of the following counts):  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 | ****

*** Incorrect number / missing attributes, -1 points for Q3

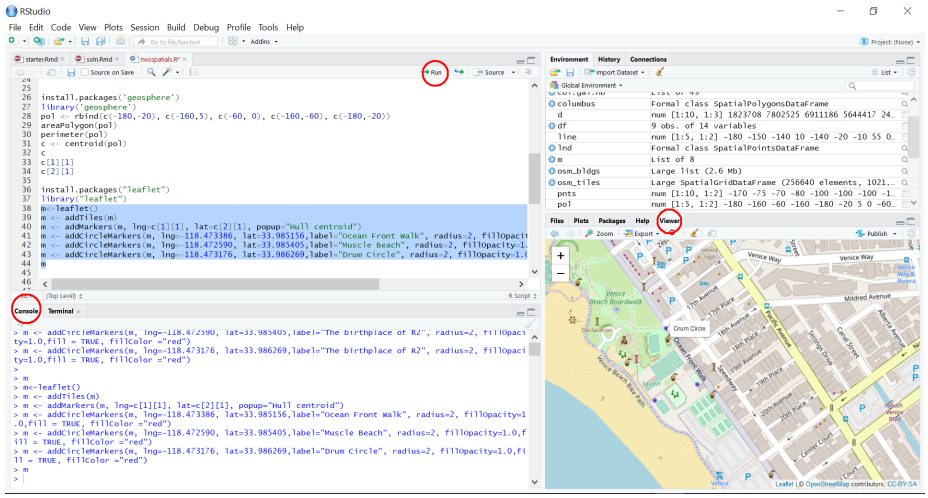
R

Q4 (0.5 point). Take a screenshot of the entire RStudio IDE that shows the code, console output and the map (sufficiently zoomed in) with the hull and centroid markers visible, for submission.

POINTS	DESCRIPTION
0.5 point	Screenshot shows code, console, sufficiently zoomed in map, with visible hull (circles) and centroid markers (dark blue flag icon) Plotting just the convex hull points is what's expected - not all the collected cords. [no point deduction] Also, no need to indicate the area - since this question is about centroid (of the hull) [no point deduction] .

.

.



*** *some students might not have USC region's convex hulls being DEN students etc, other map areas can be considered too.*

*** Missing code / missing map / missing markers, -0.25 points for Q4