Understanding Data Mining

Lesson 3 – Putting it All Together   Name_____

1. Today you will be creating your own data set. Your task is to use an online MLS© search engine to locate information about houses for sale in Wake County. We will use the data mining process to determine which characteristics about a house most significantly influence its listing price.

2. Go to www.fmrealty.com and click on the tab at the top that says "Search." Select Wake County as your area. Then, choose a maximum price ($250,000 is usually good). Next, click to search.

3. Choose ten different houses for sale. Try to choose your houses from different price ranges. Record information about your houses in the table below. Make sure to record the house prices in hundred thousands. For example, a house for sale for $149,500 would be listed as 149.5 in your table. Also, make sure to record ANNUAL homeowner's dues in the HOA column.

Price	Beds	Baths	Acres	HOA	SqFt

4. Next, get information about ten other houses from a classmate. Record the information below:

Price	Beds	Baths	Acres	HOA	SqFt

5. Create a spreadsheet in Microsoft Excel using the information you have about twenty houses. Give your columns simple names, as you will have to call on them as part of your R script.

6. Save your spreadsheet as a text file (.txt).

Understanding Data Mining

7. Open R and set your seed.

8. Change your directory to the one that holds your .txt file.

9. Upload your data.

10. Generate a linear model.

11. Run an analysis of variance on your house data. What appears to be the most significant factor? _____ The least significant factor? _____

12. Now it's time for variable selection. Run forward, backward, and stepwise selection to see which variables have a significant impact on home prices. Record the relevant variables as identified by each method:

forward: _____

backward: _____

stepwise: _____

13. Write a conclusion to relate your findings. Be sure to mention what factors appear to play a key role in a home's list price as well as those that don't. Offer your own explanations/opinions to supplement your conclusion.

_____

_____

_____

_____

_____

14. What other factors do you think could influence home prices? Why? _____

_____

_____

_____

15. Print a copy of your R script. Put your name on it, and then attach it to your lesson packet.