IOWA HOUSE PRICE CASE STUDY[1]

The Iowa house price case study is concerned with developing a model to predict the prices of houses in Iowa. Data and the analysis in Chapter 3 of Hull's book are in the files accompanying the book on

www-2.rotman.utoronto.ca/~hull

You are required to use the Python software to try and improve on the analysis in Chapter 3 by including additional features from the Original_Data.xlxs. Proceed as in the text and choose the first 1800 observations as the training set, the next 600 as the validation set, and the remainder as the test set.

In this assignment you should consider four additional features:

Part A (4pts): The first additional feature should be Lot Frontage where you should consider alternative approaches for dealing with missing observations.

- Try at least three approaches and briefly explain why you chose them.
- Fit all three regressions (Linear Regression, Lasso and Ridge) for each strategy you chose. In total you will fit at least 9 regression: 3 Imputation Strategies x (Linear Regression, Lasso, Ridge).
- Report on its performance, select one approach based on your results – this selected approach should be carried for the parts of the assignment.

Part B (2pts): Add the categorical feature **Lot Shape to the data set you used for Part A.**

- Fit all three regressions (Linear Regression, Lasso and Ridge).
- Report on its performance, by how much do these two features (Lot Shape and Lot Frontage) improve prediction. Provide supporting explanation for your results.

Part C (2pts): select two additional features to be added to the data set used in part B:

- Briefly explain why you chose them – you can choose them using qualitative reasoning/business intuition
- Fit all three regression models, report on its performance, and provide supporting explanation for your results.

Part D (2pts):

- Out of all the regressions fitted on Part A,B and C - select the regression model with the best performance that you have fitted so far.
- Randomly split data into training set, validation set, and test set – maintain the same splitting proportion 1800/600/rest respectively

[1] This case study and accompanying files were produced by FinHub, the Financial Innovation Lab at the Joseph L. Rotman School of Management, University of Toronto.

- Refit the regression model, report on its performance - are the results in line with Parts A, B and C after randomly shuffling the data? Explain your results.

Your submission should include your Python code (*.html and *.ipynb) and a short three page report outlining what you did.

Feel free to add as many tables and charts in your Appendix to support your reasoning; **these text elements will only be read and evaluated for grading if properly referenced on the report.**