Ames Housing OLS Regression Project (300 Points)

The ames_train data set contains approximately 2039 records. See the data description in the file Introduction_to_Ames_Housing_Data. This is a random selection of training data selected from the full dataset. Note, the index numbers have been randomized and the split between train and test is also random so you will not be able to match the test data with sale price values. You are to use OLS ("Linear") Regression to predict the sale price for homes in the ames_test_sfam dataset by building two models using the ames_train data.Note, the test data set is single family homes, the training data is all homes.

DELIVERABLES

- Read the report template. Your write up in PDF Format (no zip files). Your write up should have five sections. Each section should have enough detail so that I can follow your logic and someone else can replicate your work.**(150 Points)**

- A file that contains all the python code you used in your analysis. I should be able to run this file and get all the output that you got.

- A csv file,which has the scored records values from ames_test_sfam. There will be only two columns in this file: index and p_saleprice. You will be graded on how your model performs versus my model and those of other students in the class.

WRITE UP (200 POINTS)

This is a very structured assignment. Make sure you respond to each section and follow the directions. If you skip a section in your report you will receive zero points for that section.

1. **First Steps (40 points)**

Describe the ames_train data set so that I am convinced you understand it.

Use my shell code as a start to explore the data. Apply your creativity and go from there.

If you know how to do pivot tables in Excel, it is a great tool for Exploratory Data Analysis (EDA).

EDA was well established by John Tukey. He was a great advocate for it and developed much of what we do today.

Knowing your data consists of three components: (a) a data survey, (b) a data quality check, and (c) an initial exploratory data analysis.

(a) A Data Survey

- Take a broad overview of the Ames housing data set. Read over the data documentation. What data do you have, and what is it supposed to represent?
- In the linear regression component of this course you build linear regression models to predict the value of a property (single family home). Do you have the right data to properly address the problem? Are there observations in the data that should be excluded?
- What kinds of problems can you properly address given the data that you have? In particular if you were to build a regression model with the variable SalePrice as the response variable, what types of properties would you be valuing? Be careful about what you are doing here.

(b) Define the Sample Population

- When building statistical models you have to define the population of interest, and then sample from THAT population. Frequently you will not actively perform the sampling function. Instead, the data will be made available and you will have to

sample from it retrospectively, i.e. you will need to carve out the population of interest. In this assignment the objective is to be able to provide estimates of home values for 'typical' homes in Ames, Iowa. You may not be able to define what 'typical' is, but can use the data to find out what is atypical. Any values which are not atypical are then considered to be typical.
- Define the appropriate sample population for your statistical problem. Hint: You are building regression models for the response variable SalePrice. Are all properties the same? Would you want to include an apartment building in the same sample as a single family residence? Would you want to include a warehouse or a shopping center in the same sample as a single family residence? Would you want to include condominiums in the same sample as a single family residence?
- Define your sample using 'drop conditions'. Create for the drop conditions and include it in your report so that it is clear to any reader what you are excluding from the data set when defining your sample population.
The definition of your sample data should be clearly noted in your assignment report.

(c) A Data Quality Check
- In practice your data will not be 'clean'. You will need to examine your data for errors and outliers. Errors will not always show as outliers, and outliers are not necessarily errors.
- If you have a data dictionary that states the set of proper values for each field, then you will want to check your data against the data dictionary.
- If you do not have a data dictionary, then you will need to reason and explore your way to a proper data set.
Example 1: In this project you will be modeling the sales price of housing transactions. It should be obvious that none of these sales prices should be zero or negative. Observations with a zero or negative sales price should logically be considered to be errors.
Example 2: Suppose we had a 'small' number of housing transactions with a sale price over one million dollars, should we consider these sales prices to be valid? In this case these values could be valid data points, which would make them outliers, or they could be errors, such as 140,000.00 entered as 1,400,000. In either case they are not relevant data points if the objective is to model the 'typical' home price for the area. Tell me how you are going to treat outliers.

2. EDA (30 Points)

Pick ten variables from the data quality check to explore in your initial exploratory data analysis. Perform an initial exploratory data analysis. How do you perform an exploratory data analysis for continuous versus discrete (or categorical) data? Use scatterplots, scatterplot smoothers such as LOESS, and boxplots to produce graphics.
Note thatyou are particularly interested in the relationships between the response variable and the predictor variables.
Suggest you split your EDA into two sections in your report – one section for continuous variables and one section for discrete variables.

3. **BUILD MODELS (100 Points)**
Build at least four different LINEAR REGRESSION models.

The first model should be a **simple**(single prediction variable) model. Find the best single variable model.

The next model should be a multiple regression model with **two predictor variables**. Find the best two variable model.

You do not need to build more complex models for this assignment. More complex models will be the topic for hw02.

Show all of your models and the statistical significance of the input variables. Discuss the quality of fit, R squared and adjusted R squared, parsimony and anything else you can think of that might be of value to share.

Discuss the coefficients in the model you select, do they make sense? Are you keeping a variable in the model even though it is counter intuitive?

4. **SELECT MODELS (20 Points)**
Decide on the criteria for selecting the "Best Model". Will you use a metric such as Adjusted R-Square or AIC? Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model. Put the metrics in a table to display the results.

5. **WRITE MODEL FORMULA (10 Points)**
Write a mathematical formula that will show the model you selected. Explain your formula.

Make sure you include this as a section in your report. Do not expect that I will search your report to find it. This step should allow someone else to deploy your model.

The variable with the predicted saleprice should be named:

p_saleprice

SCORED DATA FILE (100 POINTS)

Use the python model that you selected. Score the data file ames_test_sfam. Overall scoring for your model is based on providing a prediction for every record in the test data. **Make sure you have not deleted any records in the test data and that none of your predictions are out of range.** Create a file that has only TWO variables for each record:

index
p_saleprice

The first variable, index, will allow me to match my grading key to your predicted value. If I cannot do this, you won't get a grade. So please include this value. The second value, p_saleprice is the predicted price for a property per your model.

Name your file yourname_410_hw01.csv.

Your predicted values will be compared against …

- A Perfect Model
- Performance of Other Students
- Predict the Average value for every home (MEAN)

If your model is not better than simply using an AVERAGE value, you will lose points.

BONUS

If you want Bonus Points, write a brief section at the top of your Write Up document and tell me what you did and how many points you are attempting. If I cannot see your Bonus work, I cannot give you credit. Bonus is difficult to grade and I don't have time to go back looking for it. If you don't tell me it's there, I cannot give you points.

The policy with Bonus is: ***All Sales are Final !***

- (10 Points) Once you select a model try something else. Are the results the same? Are there any differences?
- (?? Points) Roll the dice … think of something creative and run with it. I might give you points.

PENALTY BOX

- (Lose 10 Points) If you don't have PDF format
- (Lose 10 Points) If you don't have a GOOD Introduction
- (Lose 10 Points) If you don't have a GOOD Conclusion
- (Lose 10 Points) If you don't put your NAME in the file names of any files you hand in