Michigan Real Estate Assignment (MREA). Due January 27, 2021

This purpose of this exercise it to familiarize students with the ideas and techniques required to complete the Amadeus Data Project. In both cases, the objectives are twofold:

1. Generating causal hypotheses, and testing their validity, e.g. closing prices increase with house size (due to higher costs).

2. Predicting outcomes, e.g. closing prices.

A. **Causal Hypotheses**

At least three types of factors influence list and closing prices:

1. Property characteristics: house size, house features (number of bedrooms, bathrooms, etc), age, amount of land, garage size, and so forth.

2. Location: two homes with otherwise identical characteristics may still exhibit price differences because they are located in different locations (due to proximity to better schools, commercial districts, transportation). The dataset contains location info that may proxy for these underlying factors, e.g. county and postal codes.

3. Time: macro-level fluctuations in country and regional level economic activity can significantly affect prices. Near the end of the sample period (2000-2010), the “financial crisis” is likely to have caused a decline in real estate prices in Michigan.

Many of the aforementioned data variables are not continuous (like house size); instead they are categorical (county or postal code location). In the latter case, you should convert each category variable into a corresponding set of dummy variables (e.g. if there are ten counties in the dataset, ten binary dummy variables can be generated. Instructions on how to do this in python are available [here](https://chrisalbon.com/python/data_wrangling/pandas_convert_categorical_to_dummies/)). I recommend that you also create a set of time dummies (corresponding to each year in the dataset). Why?

**Important**: when performing your regression-based causal analysis make sure that you consider the “Nine Steps in Multiple Regression” discussed in the Week 1 Slides. Please report all of your results (not just the regression output)!

B. Predicting Outcomes

Apply a machine learning approach to prediction, using the dataset you constructed in Part A (including the dummy variables): first, use a subsample of the data to generate model estimates (the “training dataset”), and then using these estimates, determine the quality of the prediction based on the remaining data (the “test data”). Any number of data splits can be used (e.g. 50:50, 70:30, etc. Which is best?).

There are a variety of ML algorithms (see the JEP article posted on K2, Week 1 for an overview of methods used by economists). For this assignment, use the LASSO method. I have posted some python code in a separate text file on K2 (week 2).

**What is your objective here?** To find the set of variables (and data split) that best predicts closing prices and listing prices.

**Format of your report**: A pdf file that contains a full discussion of your Part A and Part B analyses.

For Part A: tables for your descriptive statistics and your various regression results. Describe/report what you discovered in following the “nine steps in multiple regression.

For Part B: the data splits, the set of right hand side variables, and the prediction results.

Finally, here are some useful links for multivariate regression and the LASSO regression method in python:

<https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b>

<https://datatofish.com/read_excel/>

<https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9>

<https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>