Digging to Data: Homework 2
Due: 23:59, --

The goal of this homework is to predict 2013 housing prices (the last year we have data available) using a variety of sources, including the price of houses in 2006 (the height of the US housing bubble). This will help you learn how to effectively use linear regression.

You will need to provide three deliverables for this assignment: 1) your code for completing this assignment, 2) a writeup detailing your steps (answering the steps highlighted in bold), and finally 3) predictions on a set of test data.

1. Download the data.

• *house_train.csv* - the training set • *house_test.csv* - the test set	• *id* - A unique ID given to each area • *zip* - The zip code associated with the area • *state* - The state the area is in • *county* - The county the area is in • *poverty* - The proportion of the population living in poverty in 2007 • *price2007*- The average home price in the area in 2007 • *price2013* - The average home price in the area in 2013

2. (40 points) Predict 2013 home prices using state information only. Answer these questions using all of the training data available (WARNING: by default, Rattle partitions data into validation and test splits).
    a. **What is the intercept? What does it correspond to?**
    b. **How do you get this information from your regression?**
    c. **Based on your regression coefficients, what states have the most and least expensive average homes?**

   d. **How do you get this information from your regression?**
   e. **What is the average price of homes in those states?**
   f. How do you get this information from your regression?
3. (10 points) Predict 2013 home prices from state and county information.
   a. What US counties have the highest and lowest regression coefficients? Why?

4. (30 points) Challenge: Build a regressor that best predicts average home values in this dataset. Upload your predictions to Kaggle (link to be provided). (IMPORTANT: **you must use your UMD e-mail to get access to this competition**). If your UMD username is JSMITH, use JSMITH_DID as your username on Kaggle. You are welcome to use any additional data you care to use so long as they are from 2007 or earlier. You must describe any additional data you use in the writeup. All students must make a submission to Kaggle; additional points available to those who do best (WARNING: make sure you submit something before the deadline, as you cannot use late days on the competition).
   a. Describe what you did to build the best predictor possible
   b. Give your best Kaggle score
   c. Give your Kaggle username

The Evaluation metric is RMSE. This rewards you for getting close to the true answer (and penalizes being very far away).

Your answer submission file should have only two columns: *id* and *price2013*. The *price2013* should reflect the prediction based on the information from that *id* in the house_test.csv file.

5. (10 points**)** Suppose you have 2 bags.
Bag #1 has 1 black ball and 2 white balls.
Bag #2 has 1 black ball and 3 white balls.
Suppose you pick a bag at random, and select a ball from that bag.
What is the probability of selecting a white ball?

6. (10 points) A soccer team wins 60% of its games when it scores the first goal, and 10% of its games when the opposing team scores first. If the team scores the first goal about 30% of the time, what fraction of the games does it win?