Stat 104 Spring 2015
Regression Project
Due 4pm May 4, 2015

Collaboration.
You must work by yourself on this project. No collaboration with anyone else (in this class or not) is permitted. You may consult with the teaching staff as the need arises.

Comment on the role of your TF
Please only consult with your TF when you run into a problem, not when you are just looking to have someone check your work. Otherwise, contacts with your TF may become excessive and this project tends to become a "joint project with your TF" rather than "your own project."

Overview
An individual "regression report" is required of all students. Your regression report is **due before 4:00 pm on Monday, May 4, 2015**. **TWO COPIES of your report must be submitted.** Submit the report as you would a homework assignment. Each student will analyze the same data set, described below, and also submit a set of predictions. Your grade will be based on the write-up, and how close your predictions are to the actual values (details below). The report is expected to be 5-10 pages in length.

Background: real estate transactions
The data set used in this project was gathered from the records of a real estate office in Springfield, home to the famous television family *The Simpsons*. A picture of the town is shown below.

To understand the data set and the project, it is necessary to understand how houses are bought and sold in Springfield and the rest of the United States.

When an individual or a family wishes to sell their house, they often choose to get help from a real estate agent. (This is true in the majority of cases but there is a substantial minority of cases in which individuals or families sell their house without using a real

estate agent.) The agent helps the family decide on a price for the house called the *list price*. The family signs a contract (called a listing agreement) with the real estate agent. The contract says that if the agent finds a buyer for the house who will pay the list price, then the family will sell the house at that price. The actual price when the house is sold is called the *sale price*. Usually the sale price is less than the list price, but not always.

An important part of the real estate agent's job is to help fix the list price. The agent looks carefully at the characteristics of the house, including its size, location, age, the amount of property that comes with the house (called the lot size), the number of bedrooms and bathrooms, and whether the house has various desirable features. These features include items like a basement, garage, and fireplace, and appliances that are sold with the house like a dishwasher, garbage disposer, and oven. In fixing the list price the agent considers the prices of similar houses that have been sold recently.

The real estate agent provides all of this information for a large book that is shared by real estate agents. Each page describes one house, including the characteristics that are important in establishing the list price, along with the list price. Then real estate agents who are helping families who want to buy a house can use this information to help guide the prospective buyers to the house that might be right for them. The data for this project was gathered from several of these books.

Background: House pricing models. The characteristics of a house strongly affect the sale price. The direction of the effect of most of these characteristics is obvious: larger houses sell for more than smaller houses, houses with more bedrooms and bathrooms sell for more than houses with fewer bedrooms than bathrooms, the presence of a garage tends to raise the sale price of a house, and so on. On the other hand, these features by no means provide a perfect prediction of the sale price of a house. In part, this is because some features are not recorded systematically by real estate agents. For example, whether there is a busy and noisy street in front of the house will matter, whether the house has been kept in good repair is important, and so on. (In the data set that you will be using, there is no information about location, which is usually quite important. For example, the location of the house determines what public school any children living there will attend, and some schools are regarded as better than other schools.) Even if there were a complete list of all the features there still would not be a perfect description of the price, because the price is also affected by who buys the house. Houses and buyers are all different. If a family looks at a house that has just been advertised for sale and that house is "just right" for them, it may well sell for the list price or even a little higher. On the other hand, if the house has been for sale for a long time and a family sees that the house will meet their needs only with some changes then it may well sell for quite a bit less than the list price.

A *hedonic pricing model* is a regression equation that relates the actual sale price of the house to its characteristics. Hedonic pricing models have a number of important uses.

(1) Real estate agents find them useful in deciding on the list price of a house.

(2) In most U.S. communities, owners of houses pay property taxes. The property tax is determined by the market value of the house. Each year the owner of the house gets a tax bill indicating what the local government has decided the house is worth, and from that how much tax must be paid. An important task for local government is to decide how much each house is worth each year. For houses that have been sold in the same year, the market value is the selling price, or quite close to the selling price. But for houses that have not been sold recently, the task is more difficult. (If local government does a poor job in estimating house prices for property taxes, this can become a political problem. House owners may be upset if they think the local government has decided their house is worth more than it actually is.) Sending someone around to look at each house each year and estimate its value is expensive. It is much cheaper to look at the record of the characteristics of each house, and then use a hedonic pricing model to estimate the value of the house. But the model must be pretty good: if the estimate is too low there is a loss of tax revenue, and if it is too high the house owner will complain.

(3) Many companies do business and widely scattered locations. For example, a large company may have offices in ten or twenty cities across the United States. If a company decides to move an employee and their family from one city to another, it is customary for the company to provide extra money if the family is moving to an area where housing prices are higher. This can be quite important -- for example, it may cost $500,000 to buy a house in some parts of California that only costs $100,000 in Springfield. Hedonic pricing models for different cities can help determine ratios of prices for similar houses in different cities. A few economic consulting firms do exactly this, and then sell the information to larger companies.

Project data

The Excel spreadsheet for this project is provided at the course home page in a file called **bldmodelhousedata_spring2015.xls**. The spreadsheet has 319 rows (of data) and 15 columns. There is a data description at the end of this assignment, describing each of the 15 columns. There is also a file containing 65 additional observations called **validationhouse_with_out_values_spring201.xls** which you will also need to download.

Assignment

Imagine that you work for a statistical consulting firm that has been asked to create and estimate a hedonic pricing model for Springfield. This will be a regression equation with house sales price as the dependent variable, and characteristics of the house as regressors. Here is a simple example, taken directly from the data set that you will be using (you can do better than this).

```
. regress salesprice baths lot_sq_ft live_area_sf garage_cars

      Source |       SS       df       MS              Number of obs =     300
-------------+------------------------------           F(  4,   295) =  122.65
       Model |  6.0482e+11     4  1.5120e+11           Prob > F      =  0.0000
    Residual |  3.6368e+11   295  1.2328e+09           R-squared     =  0.6245
-------------+------------------------------           Adj R-squared =  0.6194
       Total |  9.6850e+11   299  3.2391e+09           Root MSE      =   35111

------------------------------------------------------------------------------
  salesprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       baths |   4806.253   3584.293     1.34   0.181    -2247.771    11860.28
   lot_sq_ft |   1.822394     .57812     3.15   0.002     .6846318    2.960156
live_area_sf |   66.91846   8.659245     7.73   0.000     49.87673    83.96018
 garage_cars |   37810.33   3190.577    11.85   0.000     31531.15     44089.5
       _cons |  -16902.37   9528.855    -1.77   0.077    -35655.52    1850.781
```

Remember that the objective of the model is to predict the selling prices of houses with the same mix of characteristics shown in the data set. There are many models that could be set up. There is more to be decided here than just the variables included on the right side of the equation. You will also need to think about the functional form of the equation. (Example: Should the dependent variable be expressed in terms of dollars, in terms of logarithms, or something else? Will this decision in turn affect the functional form for the regressors?)

You will need to think about all of the technical problems that can arise in regression. For this data set multicollinearity, heteroscedasticity, and outlying observations may be problems. You will probably wish to carry out some hypothesis tests as part of your work; these will be of dubious value if the assumptions of the normal multiple linear regression model are badly violated.

The product for the project assignment consists of **two parts**:

(1) *Written repor*t. The report must be prepared using a word processor and submitted in hard copy (not email). The report itself must be no more than 10 pages, including graphs and any Stata output you wish to include. The report must present, at the end, a single model that is the final product of the work. This would be the model that a real estate agent, or a local tax authority, or a large company, would use (respectively) in the three applications discussed above.

(2) *Predictions Uploaded to Kaggle*. By the date the report is due, you must upload to Kaggle (details below) an Excel csv worksheet that has your predictions. The file will look as follows:

A	B
ID	Prediction
320	109308
321	116572
322	112316
323	95157
324	113267
325	117646
326	111169

Your ranking in the Kaggle competition will be a part of your grade. Your TF can assist you in uploading your predictions to Kaggle if necessary. Note that you are allowed up to 5 uploads per day.

The prediction contest website is [typo can't be fixed apparently] :
https://inclass.kaggle.com/c/stat-104-spring-2015-prokect

You will need to create an inclass.kaggle.com account to participate in the contest. I will send an e-mail to each student containing a clickable link with an invitation to join the competition.

For those curious, kaggle ranks how well you are doing by calculating the mean absolute error:

Mean Absolute Error = Mean |Y-Yhat|

Evaluation
The grade on the assignment will depend on three things.

1. *Clari*ty (45%). The report must clearly indicate how you went about your work, including what models were considered.

2. *Substance* (45%). The project should use the tools developed in our class in an appropriate and correct manner. The report should anticipate questions that a technically critical reader might ask. For example, if the model can predict negative sale prices for certain reasonable values of the regressors, and there is no discussion of this fact, there are problems. Similarly, if heteroscedasticity is likely to be a problem with a particular function form, then the report must indicate how this was handled.

3. *Predictive power* (10%). Using a ranking method, you will receive points based on how good your predications are according to Kaggle.

Notes:
Note that we enjoy graphs like regression diagnostic plots and scatter plots of response versus explanatory variables (at least one).

Clearly walk the reader through the analysis performed, but in a readable way without using a lot of computer output or jargon.

Your report should have a conclusion section written in "non-statistical" style that clearly summarizes the major conclusions from the final analysis and discuss these results. This section should also identify any major weaknesses of this analysis and discuss the likely impact of such weaknesses on your conclusions. The final piece of this section should discuss any policy implications of the study and possible future studies that may be needed to follow-up or confirm the current findings. Try not to get carried away with the implications.

Data Description

Variable	Description
salesprice	Selling price in dollars
lot_sq_ft	Size of lot in square feet
condition	House condition on 1-10 scale (10 best)
bsmt_sf	Square footage of basement
cac	1 if home has central air conditioning
live_area_sf	Square footage of living space
baths	Number of bathrooms
bedrooms	Number of bedrooms
fireplaces	Number of fireplaces
garage_cars	Number of
deck_sf	Square footage of deck
month_sold	Month of year sold (1=January)
age	Age of house in years
dist_fire	Distance from closest fire station
smoker	1 if previous owner was a smoker