# Kaggle Project

# Introduction

The project assignment in IS-6489 involves participating in a Kaggle competition. This particular competition, House Prices: Advanced Regression Techniques, is in the playground section, and is strictly for fun and fame, not profit. (Some official Kaggle competitions have substantial prize money at stake.) For more information, go to the competition site: www.kaggle.com/c/house-prices-advanced-regression-techniques.

This competition has already concluded. You will still be able to submit your results and receive a score but the leaderboard will not update with your ranking.

The competition consists in predicting house prices in Ames, IA. The data, described below, has already been split into 50% train and 50% test sets at the above website (with 1460 and 1459 observations, respectively). The test set contains all the predictor variables found in the train set, but is missing the target variable, SalePrice. You will use the model you develop on the train set to make predictions for the test set and then submit your predictions at Kaggle. Your predictions will be automatically evaluated on a validation data set. Your score will thus be based on the validation set performance of your model. The competition tests your ability to develop a generalizable model with low variance.

# Assignments

For this project you will work both individually and in a group to create a model of housing prices. The main assignments include two compiled R Markdown notebooks:

**Individual Kaggle Notebook**

- An individual assignment due midway through the semester.
- Develop a *parsimonious* linear model of housing prices using only 5 predictors entered additively. (In other words, for this first model do not use interactions or polynomial terms.) This limit on model terms will force you to balance predictive performance and simplicity and will ensure that you become familiar with the predictive characteristics of the variables in the data set.
- Model performance benchmark for minimum *estimated out-of-sample* $R^2$ $R^2$ is .75. (This is the model's estimated performance with new data, such as the test data.)
- Notebook should include the code you wrote for data cleaning and exploring, wrangling, modeling and cross-validation. (There should be some plots and tables.)
- Notebook should report (1) RMSE and $R^2$ $R^2$ on the train set, (2) *estimated* RMSE and $R^2$ $R^2$ on the test set (3), your Kaggle score (returned log RMSE) and rank.

**Group Kaggle Notebook**

- A group assignment due at the end of the semester. (Students may choose to work individually.)
- Develop a linear model of housing prices using *some or all of the available variables*. In this case, there are no restrictions on variables, or on combinations of variables—interactions and polynomial terms are encouraged, as appropriate. Aim for maximum predictive accuracy.
- Model performance benchmark for minimum *estimated out-of-sample* $R^2$ $R^2$ is .85. (This is the model's estimated performance with new data, such as the test data.)
- Notebook should include the code you wrote for data cleaning and exploring, wrangling, modeling and cross-validation. (There should be some plots and tables.)
- Notebook should report (1) RMSE and $R^2$ $R^2$ on the train set, (2) *estimated* RMSE and $R^2$ $R^2$ on the test set (3), your Kaggle score (returned log RMSE) and rank.

# Steps

1. Look around the Kaggle competition site. Familiarize yourself with the structure of the competition. Make sure you understand how to submit your predictions.

2. Download the train data and begin exploratory data analysis. The data set is complex; make sure you understand how the variables relate to one another as well as the structure and meaning of the missing observations. (Many of the NAs have a specific meaning in the data set that can be discerned by reading the data dictionary at Kaggle closely.) Think about how you might combine variables or create new ones.
3. After cleaning the data, develop a linear model of house prices using just 5 predictors, and submit your predictions to Kaggle. You should use a simple cross-validation method to ensure that your results will generalize well to new data. This work—and your Kaggle score— should be documented in a well-organized project notebook. This is an individual assignment due midway through the course.
4. Students will then self-assemble into project groups no larger than 3-4 to work on the group project notebook. (As noted above, you are not obligated to work in a group.)
5. Working in your group, develop a more complete linear model using as many variables as you'd like and submit your predictions to Kaggle. You should again use a simple cross-validation method to ensure that your results will generalize well to new data. Your process and results should be documented in a well-organized group notebook. This is a group assignment due at the end of the course.

# Notebook Guidelines

Your notebooks should be written using the best practices of reproducible data science and literate programming. An analysis is reproducible if it is documented in such a way that others can easily recreate and build upon it by reusing and modifying the code. Here are some rules of thumb:

- Use R Markdown. R Markdown offers a big improvement in reproducibility because it keeps your written interpretation together with your code in one document. This enables readers to see how your data analysis led to your interpretation, and to ensure that there are no mistakes.
- Comment your code. What the code is doing may not always be perfectly clear from the code itself. Commenting your code is therefore important not only for your collaborators but also for your future self, upon returning to a project.
- Use pipes. Piping syntax allows you to put individual operations on separate lines, and thereby dramatically improves the readability of your code, making collaboration and peer review easier and more effective.
- Include interpretation of plots and tables. You need to say what these mean in the context of your project; they cannot speak for themselves.

Additionally:

- Round to two decimal places any reported performance, summary statistics or coefficient estimates.
- Tables and plots should be titled.
- Use headings and subheadings within your report to organize it for your reader.
- Include a Table of Contents.
- Use bulleted lists when possible.
- Proofread! Your writing should be free of spelling errors or grammatical mistakes.
- Your notebook should be be compiled to HTML (preferred), PDF or Word.
- Warnings and lengthy output should be suppressed.
- Code should be exposed with chunk option set to `echo = T`.

See detailed expectations in the grading rubric for each assignment.

# Group Process and Grading

There are several reasons why the final stage of this project is a group assignment.

1. *Networking*. Working in a group on a hard project with people you may not already know helps you build your personal and professional network. This will surely be one of the main side benefits of your degree.

2. *Enhanced learning.* There will usually be someone in your group who is better at a particular predictive analytics task than you are, whether that is cleaning and organizing the data, engineering and selecting features for modeling, navigating the complexities of cross-validation, or iterating across models to obtain a reasonable Kaggle score. Working in a group gives you an opportunity to learn from your peers.

3. *Improved work product*. Teams are generally better at solving problems than individuals. Why? Teams have cognitive diversity— differences in perspective or information processing styles that lead to a more thorough exploration of the solution space. Simply put, difference is a strength in a team. To leverage this strength,

team members must be willing to speak their minds and contribute to all phases of the project. A recent article in the *Harvard Business Review*, “Teams Solve Problems Faster When They’re More Cognitively Diverse,” concludes: “If cognitive diversity is what we need to succeed in dealing with new, uncertain, and complex situations, we need to encourage people to reveal and deploy their different modes of thinking.”

To achieve these objectives, there must be a healthy group process, one in which individual team members are committed to the project and willing to contribute to, and cooperate on, all its phases.