**Objective**

This case study is an opportunity to showcase your knowledge of a client situation and how you would structure

and present a data-driven solution.

**Scenario**

Housing Investment Gurus (HIG) is a real-estate investment firm that buys, renovates, and resells properties in a handful of US markets. The firm is made up of experts in real estate, contracting, and marketing, and has relied over the years on the advice of its experts when making decisions on where to invest. Their industry is very competitive, so HIG would like to see if leveraging data science could give them an edge. They have hired Deloitte to help them come up with a data-driven approach for making decisions and provided recent home sales data for one of their markets, Ames, to help with constructing a proof-of-concept. In your next meeting with HIG, you will share your findings from analyzing this sample data and discuss recommendations for scaling your data science approach to other markets HIG competes in.

**Questions**

1. As part of your initial analyses, you have been asked to explore patterns in the house price data provided by HIG (train.csv). Please provide at least one visualization-supported analysis for each of the following:

* Trends in sales price over time, by year and month
* Relationship between sales price and year constructed
* Relationship between sales price and overall quality index
* Distribution of sales prices represented in the data (are they normally distributed?)

1. HIG is looking for a tool to track high value neighborhoods so that their strategy office can plan for and prioritize opportunities in high-value neighborhoods during the upcoming period. They would like to eventually develop a dashboard that can be used for monitoring and assessing “neighborhood value”, a concept that HIG has traditionally determined based on qualitative inputs from local realtors. As a data scientist, part of your responsibility is to come up with data proxies as indicators of “neighborhood value”. What proxies would you show for a “neighborhood value” view in the dashboard? And what are the top 3 neighborhoods based on your proxies?
   * Some candidate proxies could be:
     1. Features (e.g. Average/median overall quality index, GrLivArea, Garage Cars)
     2. Monthly housing price trends
     3. Number of listings
     4. How recently homes have been remodeled
2. HIG has pulled data on homes that are not currently on the market (test.csv) and wants to estimate what these homes might sell for if they go on the market in the future. They are hoping to use these sale price estimates to identify undervalued properties and would seek to buy properties that are listed far below the HIG estimated sale price.

HIG has asked Deloitte to build a model for estimating future sale price using historical sale prices (column "SalePrice" in train.csv) and characteristics of the homes that sold, and then use that model to score homes that might go on the market in the future (test.csv). If Deloitte's model is successful, HIG plans to hire Deloitte to incorporate the model into its dashboard, scale the model out to other regions, and refresh the models annually.

After constructing the model:

* + report the adjusted R-squared and RMSE of the model on the training dataset and significant features, by importance
  + identify the ten houses in test.csv with the highest predicted sale price (you can identify these houses using their index number)

In your next meeting with HIG, be prepared to discuss your approach to building the model, limitations of the current model, and opportunities for improving the model in future iterations.

**Deliverables**

* A notebook/markdown document that showcases your analyses
* A presentation of your solution that includes an explanation of your analyses and your recommendations.

**Guidelines**

* Please send your completed assignment to [grring@deloitte.com](mailto:grring@deloitte.com) by **11:59pm CST on Tuesday, November 8th.**
* You will have 20 minutes to present your findings and solution, followed by 20 minutes of Q&A.
* Budget about 6-8 hours for the entire case. Questions 1 and 2 are more purposefully exploratory and

open-ended, so make sure not to spend too much time on them.

* If you feel you are missing any information from the description, please make your best guess and make a note of it for discussion during your interview.

**Follow-up questions:**

If not addressed in the presentation, please discuss the following case-related questions:

* **Question 1**
  + Since sales prices are not normal distributed, what are your approaches to deal with this?
* **Question 2**
  + If the candidate used average price or total price as neighborhood value, ask:
    - How would you suggest HIG if there are a good number of houses listed but unsold?
  + If the candidate listed more than 3 proxies, ask:
    - If you are limited to only 3 proxies, how would you combine your choice of proxies?
* **Question 3**
  + What significant variables surprised you the most? What could be the reason for these being significant?
  + What adjustments did you have to make to the data to prepare it for modeling?
    - E.g. Up/down sampling, outlier removal, missing value imputation
  + What model algorithm did you use, and why did you choose to use that algorithm?
  + What additional data elements could be used to enhance the model further?
  + What changes will need to be made to the code and/or modeling approach to generalize this model to other geographic regions and populate the results in the client dashboard?
    - If candidate is stuck, consider these prompts:
      * What format should the data be in? (csv flat files vs pulled from a db)
      * What parts of the code can be turned into functions to enable repeatability?