Case Study 3: Understanding The Housing Market

# Classification of Assignment

* This is a Category B Assignment - A Group Assignment:
  + Your group may not receive help from anyone outside your group. All questions concerning this assignment should be addressed to your professor. It is an honor code offense to give help to other groups and individuals or receive assistance from other groups and individuals.
  + Groups are assigned by MSBA program.

# Instructions

* A dataset is provided for you labelled Case3.csv. This dataset was retrieved and mostly cleaned from an ongoing Kaggle competition <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview/description>. This dataset was edited for simplicity. A data dictionary is attached on Blackboard.

# The main goal of the assignment is to explore the factors affecting SalePrice.

* In an R script file, load the Case3.csv dataset and using the appropriate functions to inspect the data frame.
* Using the dataset Case3.csv, create a linear model that explains at least 71% of the variance in SalePrice.
* In doing so, test for the assumptions: linearity, heteroskedasticity, normality of errors, and multicollinearity.
* Transform x or y variables to combat skewness which leads to heteroskedasticity (if present in your model). Give a statement of why you chose what you did and how that affected the model.
* Delete observations that are considered strong outliers help pass the assumptions. Describe how you found those outliers and why you chose them.
* Provide your final regression equation.
* Provide the final correlation (if a log Y model was used) or adjusted R square value (if no log Y model was used) that is applicable to the model.
* Provide a summary in comments interpreting how well the model did.
* In a team RScript file, include the following:
  + A .R file that saves and opens without edits.
  + Code that runs procedurally from top to bottom without error.
  + Your team name along with individual names listed of people that contributed in comments at the start of the script
  + Data loaded properly directly from your working directory (no subfolders) with summary functions helping to understand what all is included in the dataset.
  + One final regression model with assessment as to what the statistics in the model refer to.
  + A test of the assumptions and an assessment (in comments) as to whether you think they were violated or not in your final model. An assessment about what your findings do to the integrity of the model.
  + Any other visualizations that describe the relationship or just the variables along with how those visualizations help you.
  + The results along with your implication (in comments) of why we care about the relationship describing what useful information someone can extract from what you found.
* One required R Script file per group should be submitted via blackboard by the due date listed in the system.

Note – splitting the dataset into training and testing groups is not required for this case study.

# Rubric for Paper (80%)

A (100) – This assignment is considered **exemplary**. The R script file was **clean and free from errors and code ran from start to finish**. All parts included. Assessments correct and complete.

A- (90) – This assignment is considered **well-done**. The R script file was **free from errors and code ran from start to finish**. All parts included. Assessments mostly correct, but **missing minor detail.**

B (82) - This assignment is considered **proficient**. The R script file had **minor errors** preventing it from running without edits. All parts included. Assessments mostly correct, but needed **clarity and detail**.

C (75) - This assignment is considered **acceptable**. The R script file **had errors** preventing it from running without edits. **Most** parts requested were included. Some assessments **incorrect or needed clarity and detail.**

F (60) – This assignment is considered **underwhelming**. The R script file was submitted, but **major** **detail was missing and or inaccurate.** R file Includes **some** of the necessary components. **Not all assessments correct.**

F (0) – No R file submitted.

# Group Contribution and Assessment (20%)

* Everyone in the group must contribute and write a portion of the code
* Groups will be peer evaluated through teammates. **Participation and score through teammates will be incorporated into your final average.**