**KAGGLE PROJECT (MSDS 6371)**

**H. H. Nguyen, I. Nwaogu and H. Wang**

**I. INTRODUCTION**

This is our Kaggle project for the MSDS 6371 (Statistical Foundations for Data Science). Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

To compete this project, we will use 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, to predict the final price of each home by using different methods in order to choose the best model.

**II. DATA DESCRIPTION**

**III. ANALYSIS OF QUESTION 1**

* Restatement of Problem
* Build and Fit the Model
* Checking Assumptions  
  - Residual Plots  
  - Influential point analysis (Cook’s D and Leverage)  
  - Make sure to address each assumption.
* Comparing Competing Models  
  - Adj R2  
  - Internal CV Press.
* Parameters  
  - Estimates - Interpretation  
  - Confidence Intervals.
* Conclusion  
  - A short summary of the analysis.

**IV. ANALYSIS OF QUESTION 2**

* Restatement of Problem
* Model Selection  
  Type of Selection  
  - Stepwise  
  - Forward  
  - Backward  
  - Custom
* Checking Assumptions  
  - Residual Plots  
  - Influential point analysis (Cook’s D and Leverage)  
  - Make sure to address each assumption
* Comparing Competing Models  
  - Adj R2  
  - Internal CV Press  
  - Kaggle Score
* Conclusion: A short summary of the analysis.

**V. APPENDIX**

* **SAS codes and Outputs for Analysis of Question 1**
* **SAS codes and Outputs for Analysis of Question 2**