**MSDS 6372 Project 1**

*Zackary Gill, Tej Tenmattam, Limin Zheng*

1. **Introduction**

With the large number of houses available on the market it is difficult to have a professional go through a home and come up with a reasonable price for a home. Our goal is create a few models that will allow houses to be quickly and accurately priced. To accomplish this goal, we have worked on some detailed EDA and many different modeling techniques to identify an algorithm that performs better with a train/test sets RMSE-score. Two of our models will be complex with the third being easy to explain in order to allow people to quickly see what the most important things are that relate to the SalePrice of their home.

1. **Data Description**

This dataset is from the Ames area of Idaho and contains 1460 observations with 79 explanatory variables and one response variable called “SalePrice”. Each one of these explanatory variables describes nearly every aspect of the residential homes in that area. For more information about the dataset go to Kaggle’s website (https://www.kaggle.com/c/house-prices-advanced-regression-techniques).

1. **Exploratory Analysis**

The initial examination of the data resulted in finding approximately 19 columns have missing data. We examined each of these and fixed those with logical values (EX: with Fence being NA, it is assumed that there is no fence). Next we removed columns that had factors that had problems with their levels (EX: Utilities had two levels, 1459 of the rows were of one level and the final row was of the other level). Removal of columns that had too much missing data and the consolidation of redundant columns were next. We examined the correlation plots and removed a few highly correlated ones that were describing similar attributes. At this point the data was clean so we moved on to the model building analysis.

*Transformations, Pairs plots, correlation plots, removal of columns, adding baths, put in charts here.*

1. **Objective 1**
2. **Problem Restatement**

…and the overall approach to solve it

1. **Build and Fit Models**

For our analysis we built three different models. Two of them were built with automatic selection algorithms: stepwise and lasso. The other model is custom and was designed to be easily interpretable.

Stepwise

During our analysis only the model

Lasso

**Custom**

The custom model was to be one that is we took the variables that were the most significantly correlated with L\_SalePrice and also a few sensible ones. The model consists of

1. **Checking Assumptions**

Residual Plots

Influential Points

1. **Comparing Models**

RMSE or whatever

1. **Parameter Interpretation**

Interpretation

Confidence Intervals

1. **Conclusion**

Conclusions, insights, concerns, what to do better next time?

1. **Objective 2**
2. **Goal of 2way ANOVA**

State what route you are going to take 2way ANOVA or Time series and summarize the goal.

1. **Analysis of 2way ANOVA**
2. **Conclusion/Discussion Required**

The conclusion should reprise the questions and conclusions of objective 2.

1. **Appendix**

Contains code and extra charts