**Stats Foundations Project (MSDS 6371)**

**Mingyang Yu & Allen Miller**

# Introduction

This project is focused on the Ames Housing dataset on Kaggle. We will first explore how the sales price of a home is influenced by the home’s square footage in three different neighborhoods. We will then build a model that can accurately predict a home’s sales price based off of many different features of the home.

# Data Description

The data comes from the Ames Housing dataset which describes the sale price of property in Ames, Iowa from the years 2006 to 2010. This includes 80 distinct variables, all of which help describe a property and may influence the price of a home. We looked at all the observations included in the dataset (2930) and used a mixture of these variables to build models that can aid in predicting a home’s sale price.

# Analysis 1

* 1. **Restate the problem**

The Century 21 Ames real estate company has hired us to analyze how square footage of a home effects its sales price in the three neighborhoods they sell homes in (NAmes, BrkSide, and Edward neighborhoods).

* 1. **Build and Fit**
     1. Model 1
        1. The plot {insert plot here}
           1. QQ-Plot shows non-linearity at either end of the plot

Assumption not met

* + - * 1. Observation 339 has a Cook’s Distance larger than 5.6 while observation 131 has a Cook’s Distance larger than 1.

These may be due to some unique cases

* + - * 1. Observations 169 and 190 both have standardized residuals greater than 4.
        2. Since our sample is sufficiently large, removing these four outliers should not affect our results and will be removed for farther analysis.
        3. Adjusted R-Squared = 0.3406
    1. Model 2
       - 1. Model2 = model1 run without outliers
       1. The plot {insert plot here}
          1. QQ-Plot shows near linear distribution of residuals
          2. All observations have a Cooks D less than 0.01
          3. All observations have standardized residuals between 2 and 3 which is withing 5%
          4. All assumptions are met with this model
          5. Adjusted R-squared = 0.449
    2. Model 3
       - 1. Model3 is run with neighborhood data added in.
       1. The plot {insert plot here}
          1. QQ-Plot shows near linear distribution of residuals
          2. All observations have a Cook’s D less than 0.20
          3. Most observations have standardized residuals lower than 3 it appears this is within 5%
          4. All assumptions are met with this model
          5. Adjusted R-squared = 0.5165
       2. The adjusted R-squared model is higher for model 3 than model 2 so we will use model 3.
  1. **The Analysis**
     1. Using model 3 we will create 3 models (one for each neighborhood)
        1. Base Model:
        2. BrkSide Model:
        3. Edwards Model:
        4. NAmes Model:
     2. Analyze the plot and assumptions
        1. Looking at the QQ-Plot there is little evidence to suggest a non-normal distribution
        2. There is no evidence to suggest any major outliers that will need to be accounted for as the residuals appear in a random cloud
  2. **Conclusion – Grab Confidence Intervals for the interpretations**
     1. There is sufficient evidence to suggest that Model 3 is a good fit for the data (p-value < 0.00000001).
     2. We can interpret each sub-model of model 3 as follows
        1. Given that the Neighborhood is BrkSide, it is predicted that the Sales price of the house will increase by $871.63 for every 10 square feet added to the house.
        2. Given that the Neighborhood is Edwards, it is predicted that the Sales Price of the house will increase by $706.09 for every 10 square feet added to the house.
        3. Given that the neighborhood is NAmes, it is predicted that the Sales Price of the house will increase by $500.12 for every 10 square feet added to the house.
     3. Because this an observational study we cannot draw any causal inference. However, because it appears that the data was obtained from a random sample, we can assume that observations outside of the Ames Housing dataset will follow these findings given they belong to the sampled Neighborhoods.

# Analysis 2

* 1. **Restate the Problem**
     1. Using all the neighborhoods and all the variables, we will build a model that can accurately predict the sales price of a home in Ames, Iowa between 2006 and 2010.
     2. We will first explore the Stepwise, Forward, and Backward models and use our findings to create a more accurate model.
  2. **Clean-up and selection**
     1. We will first look at each variable and convert specific variables to categorical so we can use their factors for linear regression.
     2. After transforming these variables, we will plot each one of the 80 vs the sales price of the house to look for correlation and independence. From there we will begin to assemble a list of variables with strong correlation that may be good predictors in our regression model.
     3. 31 of the 80 variables were selected for our models
        1. We used the pool area variable to create a new variable that said Yes or No for a house having a pool.
     4. After selection we decided to look closely at our selected variables that contained N/A and transform those values to useable factors such as None for quality rankings. We then replotted these variables to confirm that there was still strong correlation.
     5. After plotting the residuals, we decided to use a log transform on the sales price to help normalize or data better.
  3. **Build and Fit Models**
     1. Stepwise
        1. Using R and our selected variables we created a model using 10-fold internal cross-validation. We then trained the model using the training portion of the data set.
           1. This provided an RMSE of 0.1494448
           2. There were a few observations that fell outside of 2 when looking at the standardized residuals, however, most fell within the threshold of 2.
           3. There was a single outlier with a Cook’s D of 1.5 we chose to keep this observation
        2. We then loaded the test data set and transformed and cleaned the variables that were changed in the training set.
           1. Final Results

Kaggle Score = 0.15372

CV Press = 0.1494

Adjusted R-squared = 0.8975

* + 1. Forward Selection
       1. For Forward Selection we wrote a function to generate 5 different combinations to Cross-Validate our Model.
          1. We used a 90/10 split on our data to train the model
          2. The QQ-Plot shows most of the residuals are distributed linearly however there may be some outliers towards the beginning of the plot.
          3. Cooks D shows most of the residuals being less than the 0.003 threshold.
          4. The standardized residual plot shows mild variation in the residuals however this may just be due to random noise and we will leave these observations to not overfit the model.
       2. The forward selection is then done using AIC set to 0.15
          1. Kaggle Score = 0.15432
          2. CV Press = 0.14199
          3. Adjusted R-Squared = 0.89575
    2. Backward Selection
       1. Backwards Selection we wrote a function to generate 20 different combinations to Cross-Validate the model.
          1. We used a 95/5 split on the training and test data
          2. The QQ-Plot shows most of the residuals are distributed linearly however there may be some outliers towards the beginning of the plot.
          3. Cooks D shows most of the residuals being less than the 0.003 threshold.
          4. The standardized residual plot shows mild variation in the residuals however this may just be due to random noise and we will leave these observations to not overfit the model.
       2. The Backwards Selection is done using a p-value 0f 0.1
          1. Kaggle Score = 0.15505
          2. CV Press = 0.1306067
          3. Adjusted R-Squared = 0.89548

# Reference

# Appendix

* 1. Analysis 1
     1. Model 1