# DAT 530 Advanced Statistical Methods

Project Review: Week 3

**Reviewer:** Brandon Hosley

September 10, 2020

## Project Title:

Investment Opportunity in Ames, Iowa

## Author(s):

Jessie Wang

## Source:

NYC Data Science Academy.

## The Problem the Author(s) is Trying to Solve in the Project:

This author looks to apply data science principles to approach real estate in a growing community as an investment opportunity.

## Machine Learning (ML) Algorithm(s) used:

- Mode and Median Imputations
- Correlation Matrix
- Lasso and Ridge Regressions
- **Random Forest**

## A Brief Description of One of the ML Algorithms used:

A **Random Forest Model** is a type of decision tree designed to compensate for the standard decision tree's tendency to overfit on training data. Importance of features is determined by measuring changes in accuracy of the model when certain weights are perturbed. Larger changes in accuracy are evaluated as more important. The forest is an ensemble of constituent decision trees that each classify the data, a 'vote' of the outputs gives the final classification. Overfitting is reduced by having low correlation between the constituent decision trees.

## Metrics Used to Evaluate the ML Algorithms:

**Root Mean Square Error** is used to evaluate the author's model. The author was not explicit about which values the RMSE was compared against. It appears that it may have been on the same scale as the log-value of the homes. If this is the case, the model has very good results. If the model was intended to predict the absolute home value the RMSE is so low that it suggests some sort of mistake.