**Milestone Report**

**Define the problem and Identify the client**

Real estate in the US has been an interesting topic for not only developers or investors but also to academic researchers or data enthusiast who wants to understand more about how to predict the house price. In this project, we will focus on the real estate in Nashville where the markets trends indicated an increase of $19,000 about 7% in median home sale over the past year. The average price per square foot for this same period rose to $195, up from $180. The objective of this project was to determine the features that affects house price in Nashville using linear regression with regularization. In order to achieve this, we had to take several procedures like collecting and preparing the data set ready for modeling. Visualization for this project was essential to find the insight from different features against price. We trained the data set for modeling and test it with the tested data set. During this process, we also had to use regularization like Lasso for feature selections.

**Describe your data set, and how you cleaned/wrangled it**

In this dataset, there were 56,000+ observations with 29 variables and 31 columns. Firstly, to deal with missing value in categorical values like exterior wall and tax district, I had to find the most common value for these features in order to the replace the value against the missing value since forward filling or backward filling would not work in this case. For other numeric variables, I will use their means to replace the missing value. These methods could be done using function fillna() in pandas library. There were not any outliers presented in the data. Lastly, due to the model I used required the categorical value to be dummy variables, I had to convert the variables to binary value such 0 and 1 to different columns. I used pandas to\_dummie() function to tackle this task.