**Predict Housing Prices in Miami**

**Project Proposal**

**What data set will you use? If it is available online, include a hyperlink to it.**

Miami housing dataset - <https://www.kaggle.com/deepcontractor/miami-housing-dataset>

**Number of rows in the data set**

There are 13932 rows in the dataset

**The total number of variables you will include in the analysis**

There are 17 variables in the data set, and I will include 14 variables, including the response variable “SALE\_PRC,” for the analysis. I will not use “LATITUDE,” “LONGITUDE,” and “PARCEL NO” variables to predict housing prices in Miami.

**Number of categorical variables**

There are three categorical variables in the dataset, which are,

* avno60plus: dummy variable for airplane noise exceeding an acceptable level
* structure\_quality: quality of the structure
* month\_sold: sale month in 2016 (1 = jan)

**What data cleaning do you anticipate doing?**

* Imputing missing data if there are missing data.
* Transforming variables to reduce skew – There are 11 quantitative variables in the analysis dataset, and some of them are highly skewed. I will use log transformation to reduce skewness.

**Do you feel confident about how you will accomplish the data cleaning you anticipate doing?**

It seems the dataset is pretty clean, as I did not see any missing values or negative values. Some variables are highly skewed; therefore, transformation is necessary to reduce the skewness.

**What question(s) will you investigate using this data set?**

* The major aim of this project is to predict the house prices in Miami based on the features in the dataset.
* What are the most important features that will decide the house prices in Miami?
* Find the best model that can predict housing prices in Miami.

**What specific group would be interested in answer to your question? Why?**

The target audience of this analysis is,

* The real estate industry evaluates the property value in the future and makes educated guesses about how much profit each will make.
* The general population, who plan to buy a house to know the price range in the future, can arrange their finance well.
* Property investors to see the trend of housing prices in the Miami area.

**Which 2 data mining approaches will you use? Describe how these will allow you to answer the question(s) you chose in part c.**

I will use **Multiple Linear Regression** and **Random Forest Regressor** to predict housing values in Miami. Linear Regression attempts to model the linear relationship between the sales price and the predictors. Random Forests models can help perform implicit feature selections as they provide good indicators of the important features to predict housing prices.

Note: Since this dataset has 13K rows I might use **XGBoost** instead of Random Forest