**Final Project: Predicting House Prices**

**Business Problem:** To predict asset values by using models and some factors.

**Insights:**

1. After the analysis, I discovered that living area, median income, population, the position city-state of house and legal land area are *five most* important factors that may affect the price of a house in both two models. Among all these five factors, only city state is a categorized factor and others are numbers.
2. Also, residential building style, structure exterior finish and the age of the building are also important factors, among which residential building style, structure exterior finish are categorized factors.
3. The distribution of average housing price, legal land area, living area and house age are not in normal distribution. Among these factors, average housing price, legal land area and living area are left-skewed and the house age is distributed without certain rules. Thus, the log of average housing price, legal land area and living area will help them improve the normality while taking the log of house age won’t.
4. The PIDs of top 5 houses that the linear regression did the best predicting are 2010688003, 1812121000, 1903609000, 1810350000, 1810691000. The PIDs of top 5 houses that the random forest did the best predicting are 1810918000, 1805538000, 1812624000, 2002085000, 1812639000. The PIDs of top 5 houses that linear regression didn't predict well are 1810771000, 1901863020, 1102563000, 1901923000, 1901945010. The PIDs of top 5 houses that random forest didn't predict well are 1902770000, 2003696000, 1901775000, 2002245000, 1901539000.

**Model Performance:**

1. The RMSE value for both test part (64258.77) and train part (458357.75) of linear regression model are larger than random forest (54491.09 for both). And R-Squared values for linear regression (0.81 for test and 0.77 for train) are smaller than random forest (0.88 for both). Thus, random forest model performs better so I used this model for prediction.
2. The sequence of importance of each variable in two models are different. (ie. In linear regression model, city-state of Dorchester Center, MA has greatest importance (>40) among all the city-states. In random forest model, city-state of Jamaica Plain MA has the greatest importance (>0.025) among all the city-states.)

**Recommendations:**

1. Since the assets value has a strong relationship with living area, median income, population, the position city-state of house and legal land area, the taxes of residential single-family homes in the greater Boston area should also have a strong relationship with these factors. Thus, I suggest that we can predict assets value with random forest model with these factors and give out a reasonable price of tax based on prediction.
2. Larger living area, higher median income, denser population and larger legal land area will lead to higher housing price. And house in some certain cities like Cambridge, MA will have higher prices.
3. Tax is different from state to state so when considering the city-state factor, we may take consideration of the origin consumption tax.