**Predicting Housing Prices in Miami using multiple linear regression and KNN**

The housing market is strong in 2021, and home prices continue to climb up, and inventory remains low. The **Miami real estate market** continues to break records due to pent-up demand and low mortgage rates, fuel real estate transactions in 2021. A housing price prediction model can be a vital tool for both the seller and the buyer as it can help them make well-informed decisions. Predicting property value in the future is crucial to the competitive real estate industry and helps make educated guesses about how much profit each will make.

The dataset(Miami housing dataset in Kaggle) contains information on 13,932 single-family homes sold in Miami in 2016. Even though house prices changed significantly after the pandemic, the price decision factors remain consistent. The dataset has 17 variables, including sale price, information about the property, distance to the ocean, nearest highway, nearest subcenter, and location. I removed LATITUDE and LONGITUDE from the dataset because this analysis aims to predict house values in Miami as a city. Also, the feature information in the dataset describes the house's location; therefore, we believe the latitude and longitude are not essential variables to predict house values. The variable PARCELNO, a unique identifier for each property, is removed from the dataset.

The dataset does not contain missing values, and most variables are highly skewed, which means there are outliers and need to transform before analysis to reduce skewness. Log transformation is used for making patterns in the data more interpretable and for helping to meet the assumptions of linear regression. Log transformation transforms quantitative variables to their log value and converts categorical variables to factors for R to treat them as categorical variables.

We used multiple linear regression and K nearest neighbor regression analysis to find the best model to predict housing values in Miami. The graph of the correlation matrix (Figure 1) shows the linear model is a good fit since the predictors are not highly correlated(multi-colinearity). On the other hand, KNN handles non-linearities well; therefore, we use these two algorithms to find the best model to predict house values. With the experiment, we used the k value from 1 to 20 as the tuning parameter and found '4' as the optimal k value for the regression. In the final model (KNN with k value 4), 90% (R squared value) of the variability in the Sale price can be explained using the selected features in the houses of Miami(all predictors except month sold). Using KNN, it is possible to predict house prices in Miami reasonably.

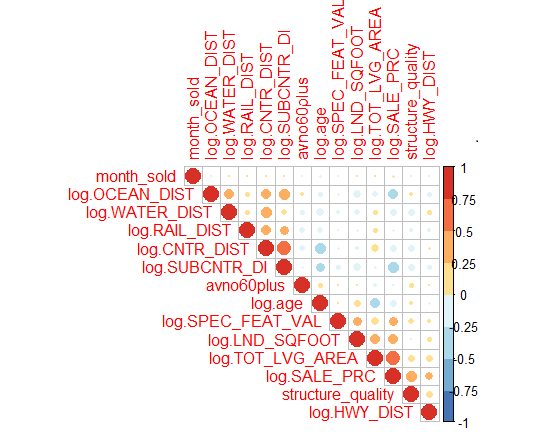


Figure 1

The most important feature of the house price prediction in Miami is the total square feet of the house(Total living area), and the second most important feature is the distance to the ocean (Figure 2). Miami's average house values are higher than the national average, and the home price has increased 20% year over year. It makes sense that the size of the home and the distance to the ocean have a more significant influence on house prices in Miami. There is a positive relationship between home size and the sale price, as shown in Figure 3. Figure 4 shows that the houses close to the ocean have a higher value than the other houses. Special feature value (e.g., swimming pool, tennis court) is the third most important feature as this indicates the facilities in the house, which added value to the sale price.

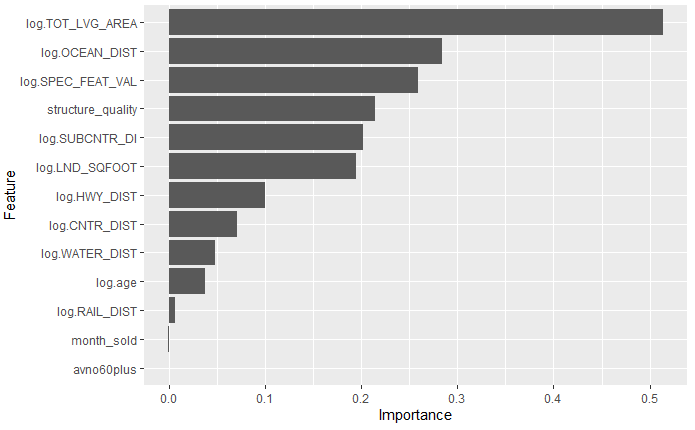


Figure 2

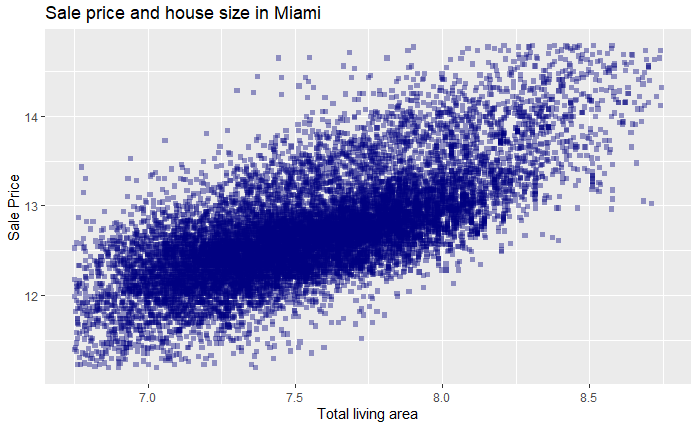


Figure 3

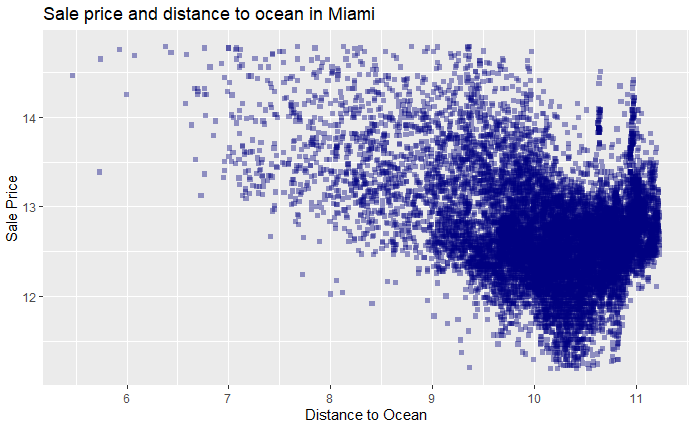


Figure 4

To improve this model, segment house prices using house types such as a luxury, median, condo, etc., will be helpful. In research from Brumer-Smith (2021), Median real estate prices have climbed steadily over the past five years. Luxury home sales are up 225% year over year as of March 2021. Condo values have increased at the slowest rate of all real estate property types and are at the most significant risk for decreased value because of oversupply. Mortgage rates are highly influential to house values; as we all know, there was a dramatic increase in house values after the pandemic. According to fortune.com, If inflation-concerned central bankers raise interest rates sooner than expected, it would translate into downward pressure on real estate prices (Lambert, 2021).

We can make this model further rigorous by using more advanced models and adding more categorical features (such as home type, school ratings, tourism, etc.) that could be useful. By adding home types such as a luxury, median, condo, etc., we can reduce the outliers in the dataset and predict the house values for each category. We did not use the "month sold" variable for our model since this dataset includes only 2016 house sale information. The prediction will be more accurate when using the last few years' data with the variable "month sold."

***References***

Brumer-Smith, L. (2021, October 15). *Miami*. Millionacres. https://www.millionacres.com/market/florida/miami/

Lambert, L. (2021, September 21). *Where home prices are going next, according to forecast models*. Fortune. https://fortune.com/2021/09/21/home-prices-forecast-models-2022-predictions/