**15/08/2022**

# HUBERT OVIE MADISE
# DATA SCIENCE / DATABASE MANAGEMENT TRAINEE
# HOME PRICE PREDICTION MODEL BUILDING PROJECT

## OVERVIEW

Home price prediction data containing parameters on home models and prices were acquired for this analysis. A `*train*` dataset made up of 1460 rows and 81 columns hosted several information and patterns for various home models and prices. A `*test*` dataset made up of 1459 rows and 80 columns hosted the same data as the `*train*` dataset but without sale price information. The goal of this project is to find patterns and relationships between home model parameters and pricing. This project seeks to build a home pricing model using relevant parameters to predict prices with utmost accuracy.

## DATA AND MODELING APPROACH

Several software tools were utilized for this analysis including SQL Server, Power BI and R Studio. Home price prediction data was acquired and imported into the `*HomePrediction*` database on SQL Server. The `*train*` dataset was then imported to Power BI for exploratory data analysis via connections with SQL Server. With the `*train*` dataset on Power BI, several dashboards were built to manually visualize and observe patterns in the dataset. Also, the Home Price Prediction Data was acquired with a `*Description*` TXT file, which served as a data dictionary and promoted better understanding of every column in the dataset. The `*Description*` file also revealed several hierarchical categorical variables which happened to be encoded in the datasets. After successful visualization and exploration of the `*train*` and `*test*` dataset, it was loaded to R Studio for further analysis and model building via connections with SQL Server.The `*HomePrediction*` database on SQL Server was also connected to R Studio as an extension of the data pipeline for further analysis and model building. This interconnectivity of software tools and datasets promoted data integrity within the data pipeline system.

Since better understanding of the data set was built during exploration, the following steps were taken for data analysis and model building:

**STEP1:** A connection with R Studio and `*HomePrediction*` database on SQL Server was created.

**STEP2:** `*train*` and `*test*` tables were extracted from the `*HomePrediction*` database on SQL Server into R Studio.

**STEP3:** Copies of these `*train*` and `*test*` tables were also created as R Dataframes to ensure data integrity throughout the modeling process.

**STEP4:** Further data exploration was carried out on the `*train*` dataset with R studio which revealed no empty rows and multiple columns with categorical variables.

**STEP5:** Columns with hierarchical categorical variables were encoded to numerical variables depending on the scale and number of ranks present on both copies of the `*tain*` and `*test*` data.

**STEP6:** Columns that fall under the numerical class were extracted for model building.

**STEP7:** A train-test split of 80:20 was executed on the `*train*` dataset. This step aimed to estimate the performance of the home price prediction model with 20% of the `*train*` dataset after building the model with 80% of the same.

**STEP8:** A Stepwise Linear Regression Model was built to predict home sale price that included 16 parameters from all numerical columns in the 80% train-test split data.
**STEP9:** The home price prediction model was first applied to the 20% train-test split data from *step7* after eliminating the pre-existing home prices.
**STEP10:** A table containing `*house_id*`, `*actual home sale price*` and `*predicted sale price*` was created and saved in CSV format for the  20% train-test split data.
**STEP11:** A graph of actual and predicted sales prices revealed high performance of model.
**STEP 12:** After confirmation of high performance from *STEP11*, the Stepwise Linear Regression Model was then applied to the `*test*` dataset.
**STEP13:** The `*test*` data was updated with predicted sales price column and saved in CSV format.

The saved CSV files were all imported to the `HomePrediction` database and then connected to Power BI for further analysis.

## RESULTS

The following was deduced from analysis and model building

1. From exploration several factors contributed to home price prediction, including the year sold - Home price increased with time and a lot of other factors could be considered when dealing with this from an economic stance.
2. Out of 16 parameters that influenced home price prediction, Overall Quality rating played the most role and had a linear relationship with price of home.
3. With an actual average home price of $176.96k for the 20% train-test split data and $177.54k average home price by the predicted model, this stepwise linear regression model can be trusted with at most 96% accuracy.

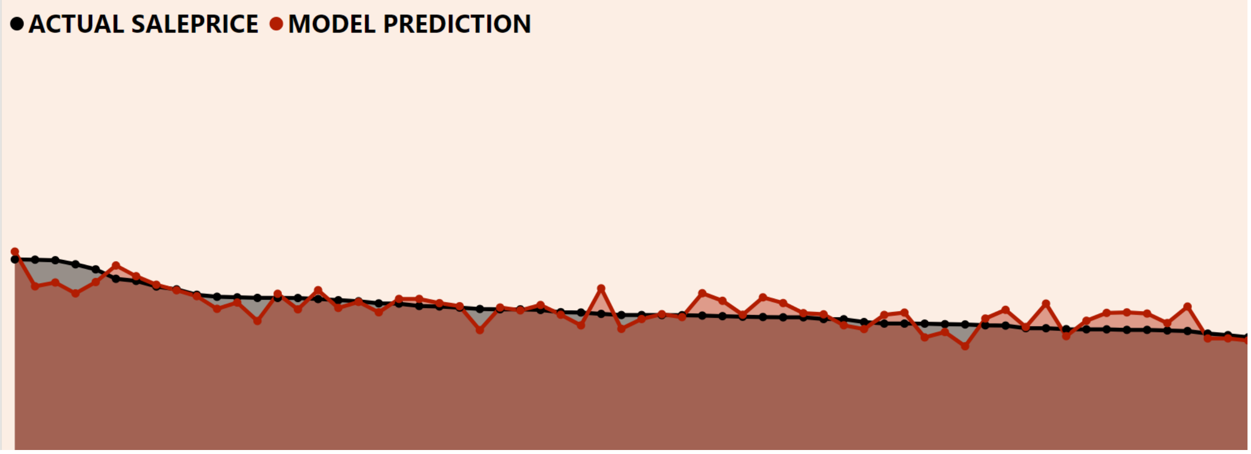

*FIGURE1: Actual Sale Price vs Model Prediction Sale Price*

## CONCLUSION AND RECOMMENDATION

Most features that contributed to sales price centered around generic physical outlook of homes. More emphasis should be placed on overall quality, External Quality and Basement Quality during home designing for Real Estate companies to maximize profit and attract a huge influx of customers during sales as they played the most role in home sale price determination.