**Github Link:** https://github.com/sagarsk2/CS584_Final_Project/tree/main

**Dataset Link: https://www.kaggle.com/datasets/gabrielemaggioni/real-estate-valuation-new-taipei-city**

## Section 1: Introduction

The goal of this project is to predict the house prices per unit area of houses in the Sindian Dist., New Taipei City, Taiwan using a combination of variables that are the house age, distance to nearest MRT station, the number of convenience stores in the living circle on foot, the geographical latitude and the geographical longitude. The data was provided by Prof. I-Cheng Yeh from the Department of Civil Engineering, Tamkang University, Taiwan.1

The project starts with an exploratory analysis of the data regarding its structure, type of variables, possible missing data and some plots to understand the relationships between the variables. Then in the analysis section of the data, three approaches are used. The first is a simple multiple linear regression model. After this, the multiple linear regression is run again but using variables using a backwards selection approach and then performing ridge regression. The third model is a random forest model. The results of all the three models are compared based on the prediction errors. The project ends with a discussion of the results and the main findings of this project are summarized using a few bullet points.

## Section 2: Exploratory Data Analysis

There are a total of 6 columns in the provided dataset. They are as follows:

1. X1: The age of the house

2. X2: The distance to the nearest MRT station

3. X3: The number of convenience stores in the living circle on foot

4. X4: The geographic coordinate, latitude

5. X5: The geographic coordinate, longitude

6. Y: The house price of unit area, to be predicted from the above variables.

The structure of the data, the pair plots and the correlation plots for the variables can be seen below:

```
<class 'pandas.core.frame.DataFrame'>
Index: 414 entries, 1 to 414
Data columns (total 6 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   house age                               414 non-null    float64
 1   distance to the nearest MRT station     414 non-null    float64
 2   number of convenience stores            414 non-null    int64
 3   latitude                                414 non-null    float64
 4   longitude                               414 non-null    float64
 5   house price of unit area                414 non-null    float64
dtypes: float64(5), int64(1)
memory usage: 22.6 KB
```

Figure 2.1 The structure of the data

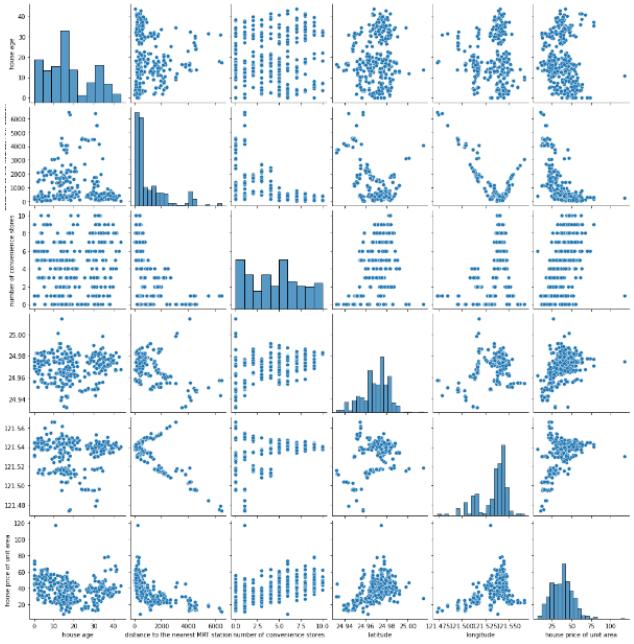

Figure 2.2: The pairplots between the variables

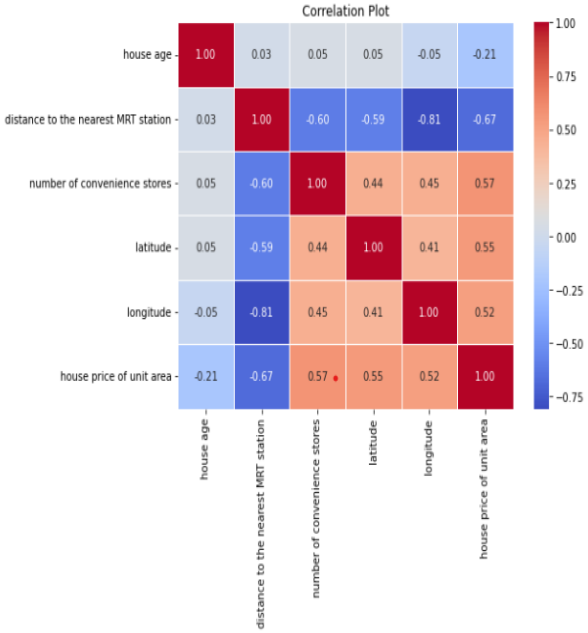


Figure 2.3: The correlation plots of the variables

| | house age | distance to the nearest MRT station | number of convenience stores | latitude | longitude | house price of unit area |
|---|---|---|---|---|---|---|
| count | 414.000000 | 414.000000 | 414.000000 | 414.000000 | 414.000000 | 414.000000 |
| mean | 17.712560 | 1083.885689 | 4.094203 | 24.969030 | 121.533361 | 37.980193 |
| std | 11.392485 | 1262.109595 | 2.945562 | 0.012410 | 0.015347 | 13.606488 |
| min | 0.000000 | 23.382840 | 0.000000 | 24.932070 | 121.473530 | 7.600000 |
| 25% | 9.025000 | 289.324800 | 1.000000 | 24.963000 | 121.528085 | 27.700000 |
| 50% | 16.100000 | 492.231300 | 4.000000 | 24.971100 | 121.538630 | 38.450000 |
| 75% | 28.150000 | 1454.279000 | 6.000000 | 24.977455 | 121.543305 | 46.600000 |
| max | 43.800000 | 6488.021000 | 10.000000 | 25.014590 | 121.566270 | 117.500000 |

Figure 2.4 Statistical summary of the columns

As can be seen, all the variables in this dataset are quantitative variables. From the structure, it can be seen that all of the data is numeric data, with 6 columns and 414 rows. Going with a 80-20 split of training and test data, 331 rows are allocated to the training data and 83 rows are allocated to the test data.

Some points to note from the summary data and correlation plots:

1. There is much more variation in the latitude values ranging from 24.93 to 25.01. This could also be due to the degree of precision of the two variables.

2. There are no missing values in the dataset.

3. The house prices are most strongly correlated with the distance to the nearest MRT station and the least with age.

4. The distance to the nearest MRT station as well as the number of convenience stores are highly correlated with the longitude which makes sense given it specifies a geographical coordinate.

5. Surprisingly these relationships are not seen with the latitude.

The above relationships were identified to be possibly important when fitting the prediction models and evaluating the conclusions. The next section explores two models that attempt to predict the house prices based on a combination of the other variables.

## Section 3: Methods

After the train-test split of the data, the two models that are explored in this project are a simple multiple linear regression model, a multiple linear regression model with backwards selection for variable selection and regularization, and a random forest model.

### Section 3.1: Multiple Linear Regression

The first fit model in this project was a simple multiple linear regression model with variables from X1 to X5 being used as predictors to predict Y. This model was chosen because it forms a good simple baseline for assessing our model performance.

Diagnostics were run on this model and the mean square error was calculated for the training and testing dataset. Moreover, the $R^2$ value was calculated for testing the accuracy of the model. Finally, a scatterplot was made between the predicted and actual Y values to get a better visual representation of the model performance.

```
Training Mean Squared Error: 85.43289947763155
Testing Mean Squared Error: 54.580945200862416
R-squared: 0.6746481382828159
```

Figure 3.1.1 The training and Test MSE and the $R^2$ value obtained from simple multiple linear regression

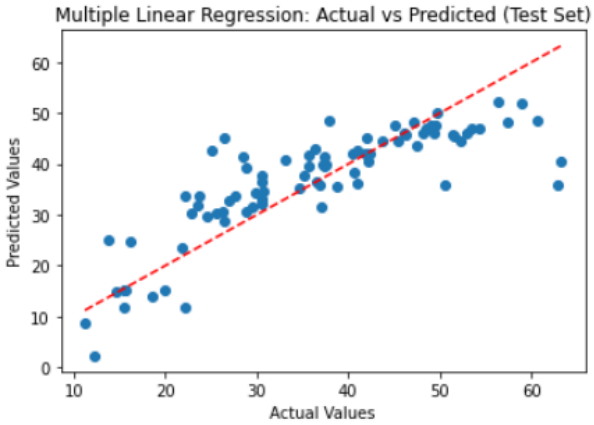


Figure 3.1.2 Predicted vs Actual Y values for the test set

As can be seen through the figures above the simple multiple linear regression model yielded an $R^2$ value of 0.67, which while greater than average, is not high enough to be treated as more than a baseline in this project. Interestingly, the test MSE was lower than the training MSE which is quite unexpected since this implies that the model performed better on the test data.

## Section 3.2: Multiple Linear Regression with Backward Variable Selection and Regularization

In order to improve on this model, feature selection using backward variable selection and regularization was done before running the regression model in order to remove redundant features and improve on the accuracy.

```
Dropped Features: ['longitude']
Training Mean Squared Error: 303.99906456602287
Testing Mean Squared Error: 52.76215401172627
R-squared: 0.6465548188527696
```

```
C:\Users\sagar\miniconda3\lib\site-packages\sklearn\utils\val
removed in a future version. Check `isinstance(dtype, pd.Spar
  if not hasattr(array, "sparse") and array.dtypes.apply(is_s
C:\Users\sagar\miniconda3\lib\site-packages\sklearn\utils\val
removed in a future version. Check `isinstance(dtype, pd.Spar
  if not hasattr(array, "sparse") and array.dtypes.apply(is_s
```

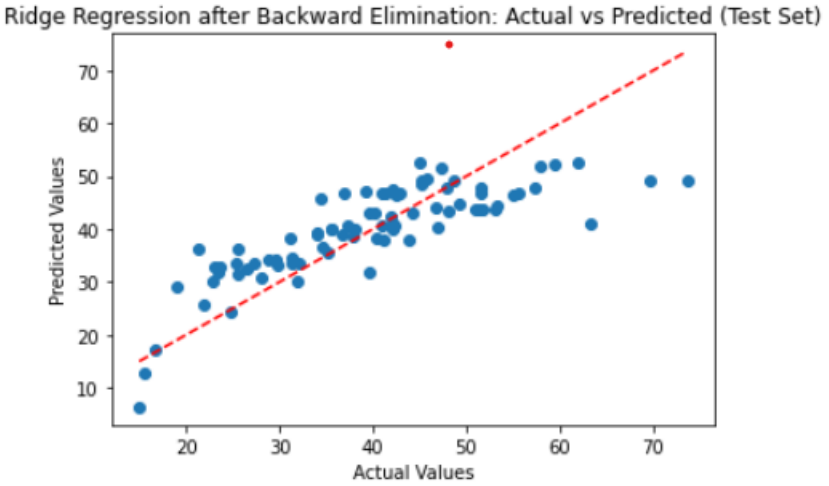


Figure 3.2.1 Training and Test MSE, $R^2$ and the predicted vs actual plot for the test set

As can be seen in this figure, despite dropping the redundant feature of longitude, the model performed nearly the same as the simple multiple linear regression model with an $R^2$ value of 0.65, which is, in fact lower than just the simple multiple linear regression model.

This is why we finally use a more complex model which is the random forest model to hopefully increase the model accuracy and improve house price predictions.

## Section 3.3 Random Forest

Finally, a random forest model was initialized and run with the results being as follows:

```
Training Mean Squared Error: 90.90560177285263
Testing Mean Squared Error: 34.60981792762854
R-squared: 0.7936941426168487
```

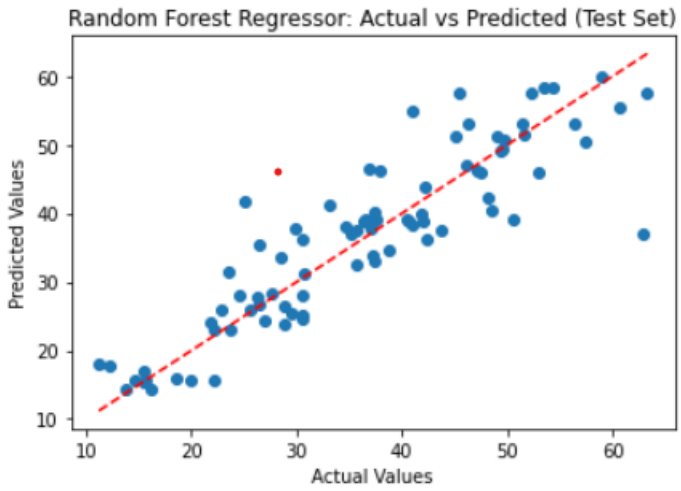


Figure 3.3.1 Training and Test MSE, $R^2$ and the predicted vs actual plot for the test set

As can be seen in this figure, the random forest model performed considerably better on both the training and test sets with a lower MSE for both and a higher $R^2$ value of almost 0.8 which is much greater than average and can definitely be applied to further test data to predict the house prices.

## Section 4: Conclusion

1. Purely in terms of results alone ($R^2$ and MSE) the Random forests model vastly outperformed the simple and adapted multiple linear regression model and therefore would be my preferred choice when faced with more data from the same environment.
2. However, there were certain assumptions for both models, especially for the multiple linear regression model that the dataset does not follow, accounting for the relatively poor performance of the model, making further testing necessary before the conclusion of the "correctness" of the models.
3. Given only about 400 values, definitely a greater sample size should prove useful in further fine tuning the parameters of the model and to make them more accurate.
4. More models such as gradient boost and regressive splines can be explored.
5. Lastly, we should collect data from more cities and sources to make the model more extensible to data beyond just that of the Sindian Dist.