Introduction

On Dec 2021, the Government implemented a package of property cooling measures in the HDB resale market as an effort to mitigate the rapidly rising prices backed by the ever-growing demand for housing in Singapore. Faced with the turbulent market forces and glooming uncertainty amongst homebuyers, this research aims to discover the key drivers of HDB resale prices and analyse their economic significance in the market.

The possible factors which have been identified as good signals to the prices are as follows:

Variable	Contextual Description
Floor Area	- Measures the area covered by the flat in square metres - Taken into account when considering the size of families
Max Floor Level	- Number of floors in the building belonging to the flat - Preference for high rise flats
Remaining Lease	- Term of ownership of the flat - Date of transaction – Lease commencement
Distance to CBD	- Proximity to Singapore's Central Business District - Key consideration for majority of the workforce
Mature Towns	- Termed by HDB as an indicator of early establishment - Preferred as they tend to be more developed with better access to many facilities
Distance to Nearest MRT	- Taken into account by travellers with MRT as their main form of transport

Discussion of Dataset & Key Descriptive Statistics

The dataset used for this research consists of 6000 transactions of HDB resale prices in 2021, taken from Data.gov.sg and managed by the Housing and Development Board. Geospatial information relating to the accessibility and frequency of amenities were sourced from OneMap.gov.sg. The outcome variable being the resale price has been normalized by dividing the values by 1000 so as to handle the numbers with more ease. Finally, the train/test set split is 2:1 where the train set constitutes 4000 observations, and the remaining 2000 observations belong to the test set.

The correlation matrix in Fig 1.1 shows that the floor area has the strongest pairwise correlation with resale prices, reflecting its influence in signalling price differences. Also, both distance related predictors have a negative correlation with resale prices, which can be explained by the higher demand for flats associated with shorter distances between key work locations and means of transport.

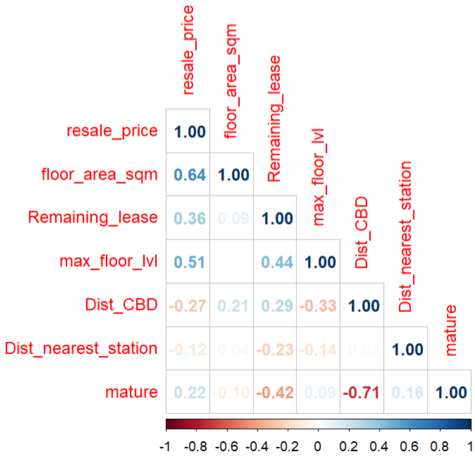


Figure 1.1 Correlation Matrix of above-mentioned predictors and resale prices

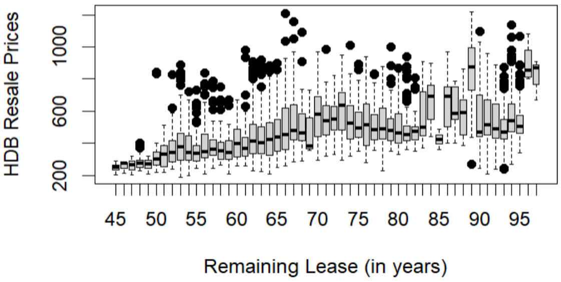


Figure 1.3 Distribution of resale prices against remaining lease

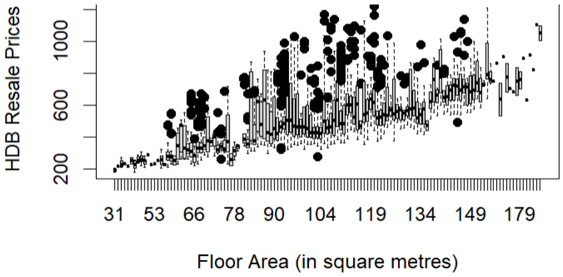


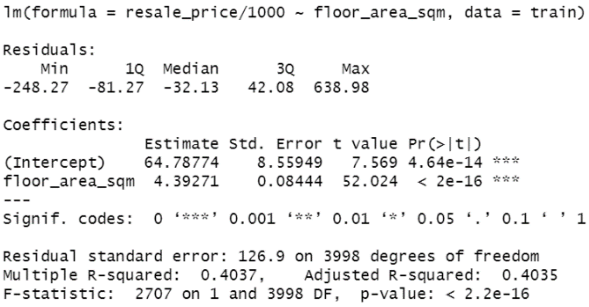
Figure 1.2 Distribution of resale prices against floor area

Based on the box plots in Fig 1.2 and Fig 1.3, a significant number of outliers can be observed across the range of the floor area and the remaining lease. This goes to show that despite both predictor variables being strong signals of resale prices, there are still other pertinent predictors which can help to explain the high variance in prices that has not been accounted for yet. The next section will analyse how to utilise these predictors to model out of sample resale prices which can aid in future projections.

Fitting of Predictive Models on Train Set

Simple Linear Regression Model

Starting with one of the more interpretable predictive models, the simple linear model is fitted on the train set with the floor area as the single predictor variable, attributed to its highest pairwise correlation with the resale prices which is supported by the high t value as seen from Fig 1.4 as well. However, the residual standard error (RSE) of 126.9 implies that the fitted values deviate from the actual prices by about $127 000 on average and the R-squared value of 0.4037 indicates that the floor area has only accounted for about 40% of the variations in resale prices overall.



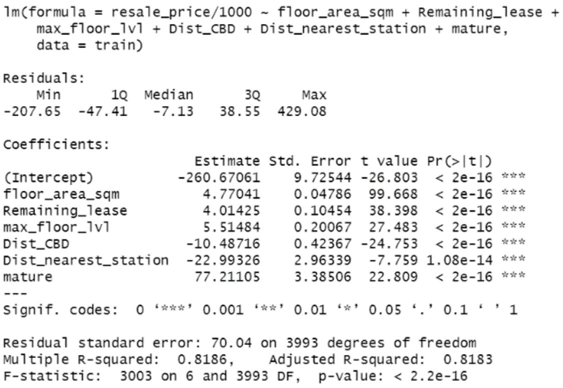
```
lm(formula = resale_price/1000 ~ floor_area_sqm, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-248.27  -81.27  -32.13   42.08  638.98 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    64.78774    8.55949   7.569 4.64e-14 ***
floor_area_sqm  4.39271    0.08444  52.024  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 126.9 on 3998 degrees of freedom
Multiple R-squared:  0.4037,	Adjusted R-squared:  0.4035 
F-statistic:  2707 on 1 and 3998 DF,  p-value: < 2.2e-16
```

Figure 1.4 Summary statistics of simple linear regression model



```
lm(formula = resale_price/1000 ~ floor_area_sqm + Remaining_lease + 
    max_floor_lvl + Dist_CBD + Dist_nearest_station + mature, 
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-207.65  -47.41   -7.13   38.55  429.08 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -260.67061    9.72544 -26.803  < 2e-16 ***
floor_area_sqm         4.77041    0.04786  99.668  < 2e-16 ***
Remaining_lease        4.01425    0.10454  38.398  < 2e-16 ***
max_floor_lvl          5.51484    0.20067  27.483  < 2e-16 ***
Dist_CBD             -10.48716    0.42367 -24.753  < 2e-16 ***
Dist_nearest_station -22.99326    2.96339  -7.759 1.08e-14 ***
mature                77.21105    3.38506  22.809  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70.04 on 3993 degrees of freedom
Multiple R-squared:  0.8186,	Adjusted R-squared:  0.8183 
F-statistic:  3003 on 6 and 3993 DF,  p-value: < 2.2e-16
```

Figure 1.5 Summary statistics of multiple linear regression model

Multiple Linear Regression Model

The earlier simple regression model is improved upon by including a few other key predictors which were discussed above. Once again, the variables have high t values which supports their relationship with the resale prices. Based on Fig 1.5, the RSE has decreased to 70.04. Due to the inclusion of multiple independent variables, the adjusted R-squared value gives a more accurate evaluation of the model’s performance. It has also increased to 0.8183. Overall, the better fitting of the multiple linear regression model on the train set can be noted here.

Non-Parametric Models: K-nearest Neighbors Regression Model and Tree regression model

Moving on to a non-parametric approach, there is no model being fit on the training data this time. Instead, optimal point forecasts on the resale prices are made by conditioning these forecasts on all the available predictors. The choice of K for the KNN model will be 2 as shown by Fig 1.6 and Fig 1.7 depicts the tree model fitted on the training data. Also, the gaussian kernel function is used instead of the rectangular function so as to improve accuracy by placing more weights on closer neighbors.

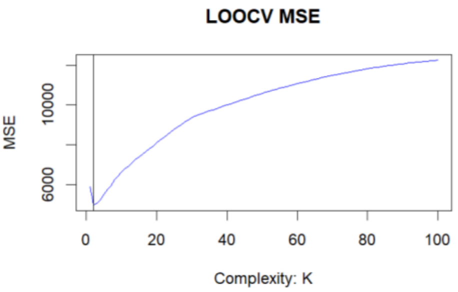


Figure 1.6 Value of K which produces minimum loss

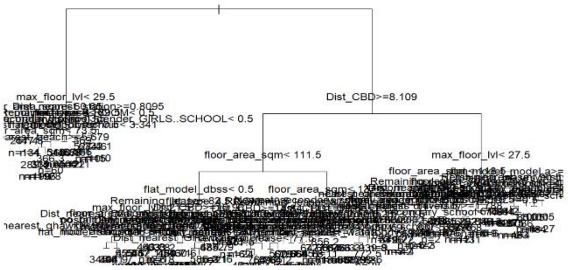


Figure 1.7 Regression tree fitted on train set

Principal component regression model

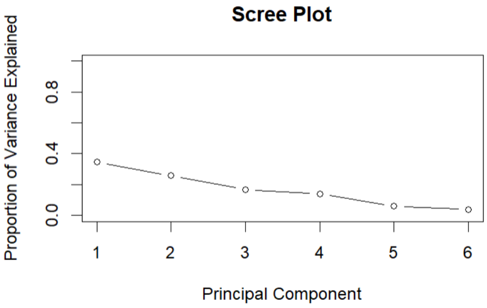


Figure 1.8 Scree plot of components (elbow point at 4th component)

Finally, this model involves the transformations of the above-mentioned predictors into principal components by having their variance modelled by these components, and the model has been fitted onto the training data. The choice of number of components will be 4 as depicted by Fig 1.8 since they explain majority of the variance amongst the predictors.

Evaluation of Models' Performance on Test Set

Model	Simple linear regression	Multiple linear regression	K-nearest neighbors	Tree Regression	Principal Component Regression
MSE	16309.84	4998.726	5331.827	3336.901	5483.754
RMSE	127.71	70.70167	73.01936	57.76592	74.05237

The simple model has shown the weakest performance as expected due to its low complexity and underfitting. The potential of nonparametric models performing better is evident here, attributable to their ability to fit flexible nonlinear functional forms while avoiding overfitting. The tree model has outperformed others in this case, which can be explained by its distinct ability to carry out automatic interaction detection and variable selection as it scales up to fit large sets of data.

Conclusion

Overall, many factors need to be considered when implementing a model to perform out of sample predictions. In this context, the PCR model might not actually be as appropriate since the variables influencing HDB resale prices are diverse in nature thus the model might exclude a key predictor when capturing their accumulated variance. Returning to the context, the property cooling measures implemented in Dec 2021 will certainly distort market transactions and equilibrium prices which wouldn't be captured by the models. Therefore, the real challenge of predictive modelling lies in achieving an interpretable model that can explain the choice of its parameters and the reasons for its anomalous predictions.