Introduction

In recent years, housing has become a more pertinent issue to Singaporeans, as we often see many headlines in the news of million-dollar HDBs being sold on the resale market, as well as numerous peoples' inability to acquire a BTO. The government has also in turn enacted certain cooling measures on the property market, the most familiar of which would be the additional buyer stamp duty (ABSD). In this study, we aim to investigate if there are any potential key predictors of HDB resale prices on the secondary market. We begin by exploring the dataset, our methodology, the results and analysis, limitations and future work, and finally conclusions.

Data

We use a fragment of a dataset of HDB resale prices from 2021, containing 6000 observations of 230 variables. We observe that some of these variables consist of zeroes throughout all observations, and hence remove them, leaving us with 213 variables. We opted to use a split of 5000 observations for training/validation, and the remaining 1000 observations for testing.

Methodology

We begin by identifying variables of interest. For this, we opted to use the following: months, some numeric discrete and continuous variables (e.g. remaining lease, floor area, number of MRT nearby, etc.), rooms sold, rooms rented, flat type, nearest MRT line, amenities in close proximity, and mature vs non-mature estates. We thus eliminate the other variables, leaving us with 88 variables, a drastic initial reduction in dimensionality. From the remaining variables, we combine some into factor variables, so as to avoid the problem of dummy variable/collinearity. These will be the months, towns and flat type variables. Next, we instantiate 3 different instances of the data set, one with resale price unchanged, one with resale price standardized, and one with resale price normalized, to observe which type of scaling (if any) returns better or more interpretable results. We then run the following models:

Correlation matrix and linear regression on all the variables for the full dataset (1)

Kernel Density Estimation on all our variables of interest (2)

Regression Tree on all variables except months and towns separately (3)

Principal Component Analysis and Principal Component Regression on numeric variables of interest (4)

K-Means on numeric variable, plotted by resale price and individual variable of interest (5)

After analysing these, we then run the following final models, using variables that we have identified as significant:

Linear regression (6)

K-Nearest Neighbours (7)

Our rationale for these models are that linear regression is an logical choice for such regression problems, due to its simplicity and for use as a benchmark, and K-nearest neighbours has been reported to perform well in datasets with high signal-to-noise ratio, which we believe to be the case in this dataset.

Results & Analysis

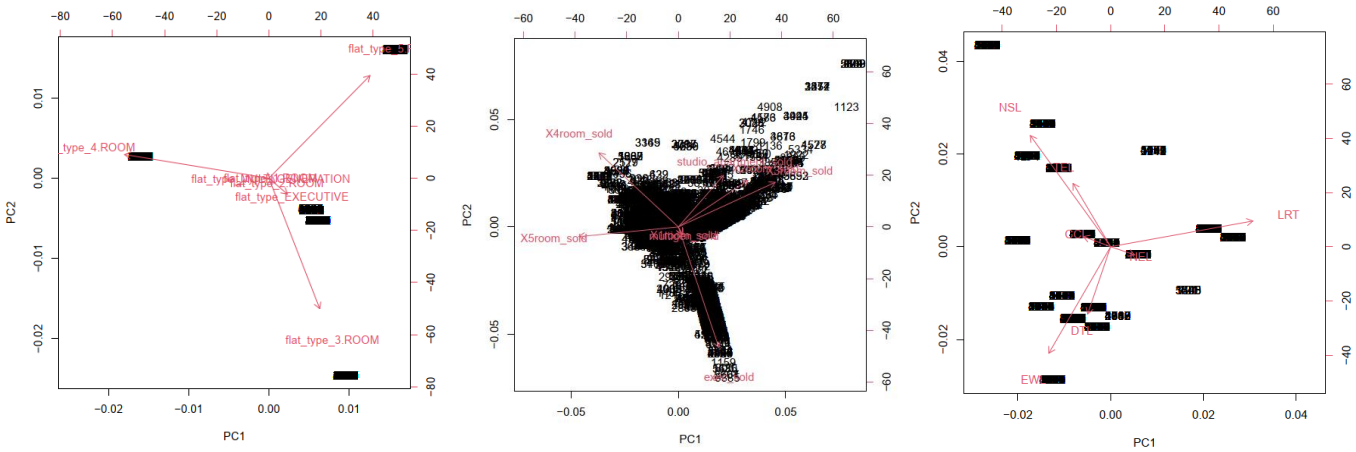
Diagrams are included in-text while tables are included at the end of this section.

From the summary of the datasets, we can see that most of the data are of rather small magnitude except the resale price, hence the decision to compare both scaled and unscaled data.

From (1), the correlation matrix returns 4 pairs of variables with positive correlations (0.8≤corr<1) at indexes 3094, 3786, 4956, 5173, 5338, 5428, 6973, 8224, which corresponds to the pairwise combinations of (nearest_ghawker_no_of_mkt_produce_stalls, nearest_ghawker_no_of_stalls), (exec_sold, flat_type_EXECUTIVE), (floor_area_sqm, flattype), and (Dist_nearest_GHawker, LRT). These may be variables we wish to examine more closely in the subsequent analyses. The significance of many of the variables in the linear model also indicate that many are contributive to the resale price of HDBs, and further dimensionality reduction has to be done.

From (2), we are able to observe the density estimations of each variable of interest. Based on (3), we notice that the most important variables appear to be “floor_area_sqm”, “Dist_CBD”, and “max_floor_lvl”, “Remaining_lease”, “flattype”. In particular, flat type with a threshold value of 4 (corresponding to a 4-room flat), appears to be the most important out of these variables. None of these are normally distributed from the KDE in (2). Flat type, as a strictly categorical variable, cannot be properly evaluated. Distance from CBD and remaining lease is heavily left skewed, while floor area and maximum floor level are heavily right skewed, which might explain the significance of these variables in determining resale price (as scarcity drives prices up). [We are unable to attach the tree images within this report as they are quite large.] The errors from the tree prediction are given in Table 1.1.

From (4), we observe that the loading value plots are not particularly useful, as there is firstly no recognizable trend or relationship between the loading value plots, and secondly the number of principal components (PCs) required to attain a relatively low threshold of 0.8 of variance explained is quite high, making the plots difficult to obtain information from. The biplots provided more insights in certain cases. For example, for flat type, we noticed that the 3, 4, and 5 room flats influenced the first 2 PCs more than the others (Diagram 1.1), which might indicate they are important predictors of HSB resale price. Interestingly, in the biplot of number of rooms sold (Diagram 1.2), we notice executive HDBs sold as a major contributor (while in flat type executive units were not a major contributor). This could imply that the selling of executive HDB units indirectly drives the resale price of other non-executive HDB flats up. Next, regarding maturity of the estate the HDB is in, we noticed it was one of the major contributors in the biplot of numeric variables. Lastly, regarding MRT/LRT biplot (Diagram 1.3), we observed that the biggest contributors were LRT, NSL and EWL of roughly equal magnitude, with a second set of contributors of smaller magnitude of TEL, DTL.



Diagrams 1.1, 1.2, 1.3, from left to right.

From the errors we get in PCR (Table 1.2), we observe they are greater by about an order of magnitude compared to the errors in (3). This seems to concur with our analysis in (1), as the variables are not highly correlated with each other, rather they have weak correlations and independence and hence PCA/PCR does not work well.

The results of (5) mostly confirm our previous analyses in (1), (2), (3), (4). Of interest is the total number of units in the HDB block, as we observe that within a certain range, HDBs seem to fetch a slightly higher resale price on average compared to the other clusters.

Based on the magnitude of the errors shown in tables 1.1 and 1.2, we decide to use the standardized data for final testing and evaluation, as it provides errors of sensible magnitude. We also narrow down our scope of variables to flat type, mature, executive sold, LRT, NSL, EWL, TEL, DTL, floor area , CBD distance, max floor level, total units, and remaining lease.

For (6), we first run a simple linear regression on the variables, and plot the residuals vs fitted graph, which indicates the possibility of some non-linear transformations required of the variables. The MSE is 0.1665874. However, even after trying various power and log transforms, we are not able to achieve homoscedasticity of the residuals. It may be that the relation between resale price and the variables are not linear at all (Diagram 2.1).

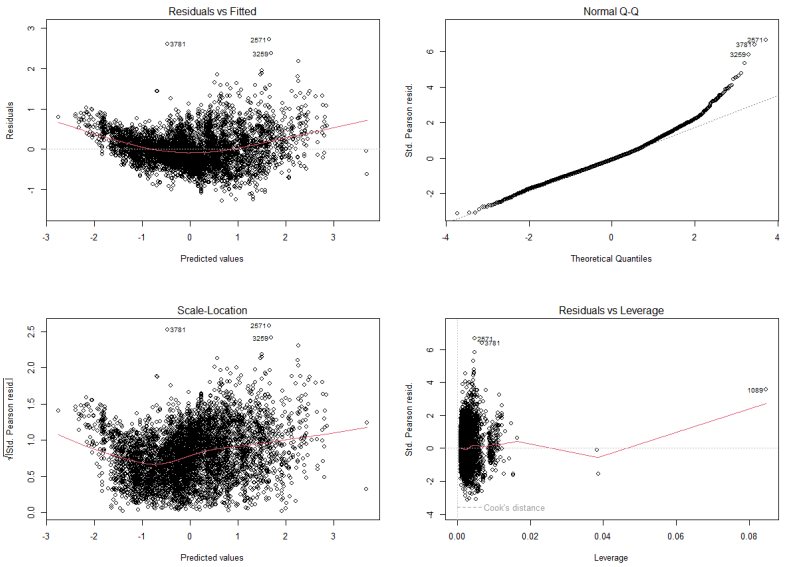


Diagram 2.1

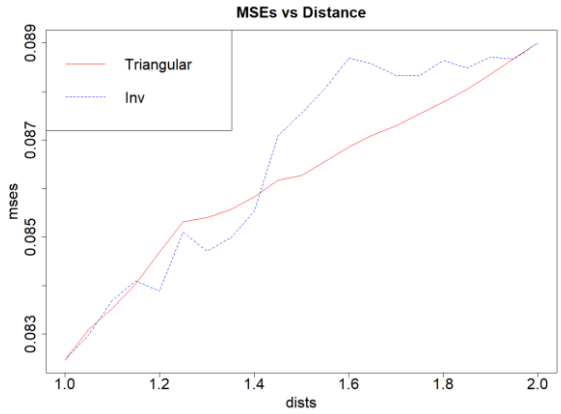


Diagram 2.2

For (7), we experimented by varying the KNN model's type of kernel, and scaling of predictors. The results are shown in Table 2.1 and 2.2, including the optimal number for K, kbest. We notice that scaled predictors return lower MSEs, and the best performing kernels are Triangular and Inv. We then use these 2 models to vary the Minkowski distance parameter, from 1 to 2 at intervals of 0.05. For both, the optimal distance parameter was at 1, and triangular kernel performed slightly better than Inv kernel, with a MSE of 0.08246549 to 0.08273608 (Diagram 2.2).

Lastly we use our best model (KNN, triangular kernel, scaled predictors, distance parameter 1) to predict on a HDB instance taken from https://services2.hdb.gov.sg/web/fi10/emap.html. For a 5 room HDB flat at Blk 101 Bishan St 12 S(570101), sold in Oct 2022, the model predicts an (unscaled) price of 813,887.8, which is 855,396.1 when corrected for inflation. This is relatively close to the actual price of 880,000 (error of 2.8%).

Tables attached below:

Errors	Month	Town	Flat Type
Unscaled	3560171409	3658485967	3409746105
Standardized	0.1337268	0.1379654	0.1280765
Normalized	0.003291579	0.003399886	0.003152502

Table 1.1

Errors	contvars	noroomsold	noroomrent	flattype	amenities	MRT
Unscaled	5903752906	17804022370	25974952723	14215121263	25868899042	25324738811
Standardized	0.2217562	0.6687529	0.9756686	0.533947	0.971685	0.9512453
Normalized	0.005458351	0.01646082	0.0240153	0.01314268	0.02391725	0.02341414

Table 1.2

Error	Rectangular	Gaussian	Optimal	Triangular	Epanechnikov	Inv	Cos
Scale=TRUE	0.1022258	0.09571128	0.0917413	0.08900196	0.09336586	0.09069561	0.09153443
kbest	2	4	6	7	4	7	5

Table 2.1

Error	Rectangular	Gaussian	Optimal	Triangular	Epanechnikov	Inv	Cos
Scale=FALSE	0.1444978	0.1311119	0.1239776	0.1226714	0.1272815	0.1185397	0.1247645
kbest	2	3	5	7	4	6	5

Table 2.2

Limitations and Future Work

There were variables that could have been explored but were not in this study. In particular, the flat model variable (in this study we only explored flat type), as well as the storey number of the HDB flat. We also noticed that resale price did exhibit large variations in both mean and standard deviation across the different HDB towns (Diagram 3.1). This should also be another factor to take into account if we have access to better geographical data.

We also did not have access to price data over the span of a year or more, and thus were unable to conduct time series analysis or test for seasonal patterns in resale price using models such as ARIMA or GARCH.

Euclidean distance may also not be the most appropriate distance function, and it may be beneficial to weight the variable distances based on perceived importance to buyers and sellers of HDB.

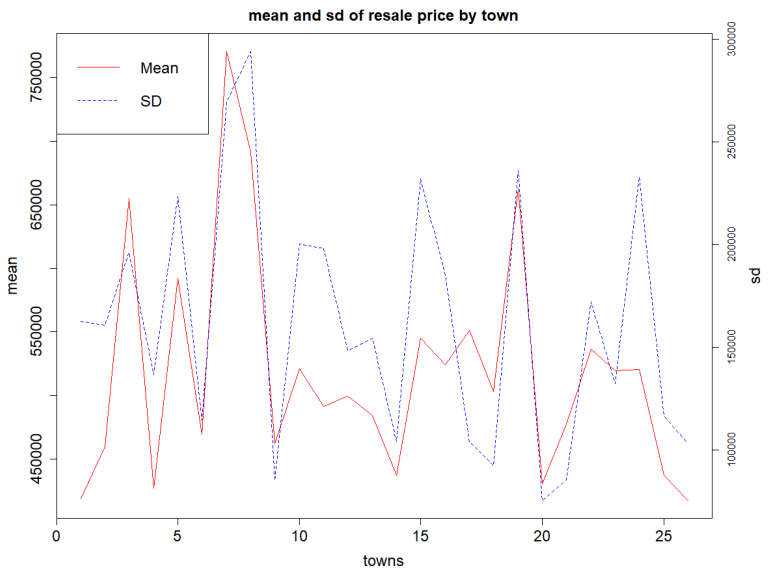


Diagram 3.1

Conclusion

In this project, we have used several methods over two main steps, in order to produce a regression model to predict HDB resale prices. In the first step we conducted mainly dimensionality reduction, by eliminating less relevant variables, and focusing on variables with more direct and logical interpretability. In the second step we trained a linear regression and a KNN model using the selection of variables determined in the first step, of which the KNN model performed better. We conducted further fine-tuning of the KNN model, before using it to predict on a random instance of a HDB resale outside the train and test sets. The price predicted was reasonably accurate with a small margin of error. We have thus determined some key predictors of HDB resale prices, and also suggested possible limitations and future works that can be conducted.