HDB Flat Resale Price Analysis

Zheng Jiaxin

A0266129Y

Summary

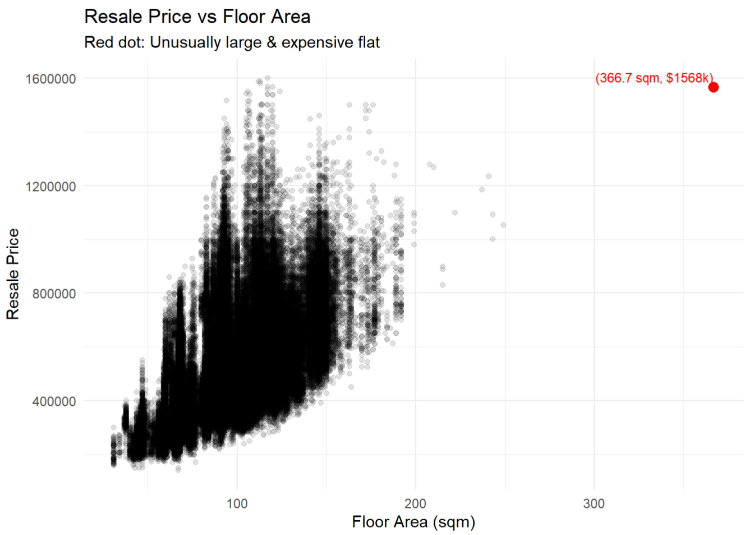
The current Residential Property Market in Singapore is experiencing a boom. To understand the market, I chose the HDB flat dataset (month, flat_type, storey_median, floor_area_sqm, flat_model_group, remaining_lease_years, region, resale_price) from 2017 to 2025 to investigate the increase in resale price. For further analysis, I built a linear regression model, natural splines, smoothing splines, regression tree, boosting regression tree, XGBoost, and multilayer neural network to see which flat type I can predict the resale price most accurately. Among all models, XGBoost achieves the best performance. I also found that while all variables that are used for modelling are strong predictors, **floor_area_sqm** turns out to be the most influential.

Data Description and Cleaning

Public housing (built by HDB) is a major component of the Singapore Residential Property Market (Phang & Wong, 1997). The dataset over Flat Resale Prices therefore is chosen considering its representation of this unique characteristic. Given the time frame of the dataset is less than 8 years, I decided to use the entire dataset for analysis. After correlation analysis, the column **lease_commence_date** is dropped due to its high correlation (0.99) with **remaining_lease**. Besides, **town** and **flat_model** is regrouped to reduce categories, and **storey_range** is converted to numeric form as **storey_median**. At the end, I had also dropped columns like block and street name which do not add meaningful locational information. The final cleaned dataset has 202747 entries with 8 columns.

Exploratory Data Analysis

Figure 1: Resale Price vs Floor Area



With figure 1, I found a strong positive correlation between resale price and floor area, which is within expectation for a housing market. Yet, what truly captured me was the red dot that is completely away from the population. This data point is found to be of flat type “3 ROOM” that was transacted very recently (2024 July). Although it seems like an invalid data point, considering the extreme floor area this data point holds, I consider it a valid yet high-leverage data point that shall be included in the analysis as part of the growing housing market. Although valid, this extreme-value flat could influence regression estimates, hence scaling was applied to ensure model robustness.

With Figure 2, I found that after the outbreak of COVID in 2020, the resale price started stably increasing. It is out of expectation given the covid has driven economic downturn for at least a year globally, but there is no noticeable decline, only minor fluctuations. Despite the growth in 2020, figure 3 confirms a continuous growth in resale price even after 2020. This is probably due to a limited supply of flats that meet MOP (minimum occupation period) (Wei, 2025).

Figure 2: Average Resale Price Over Time

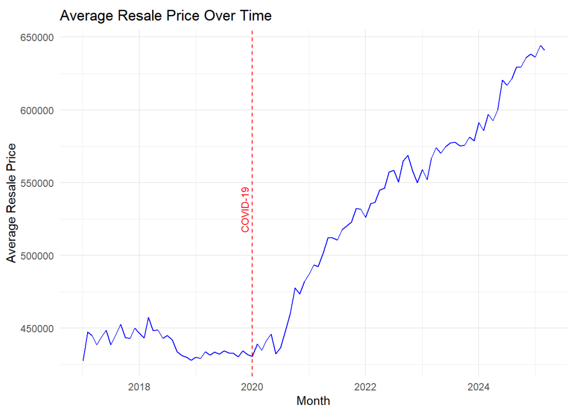
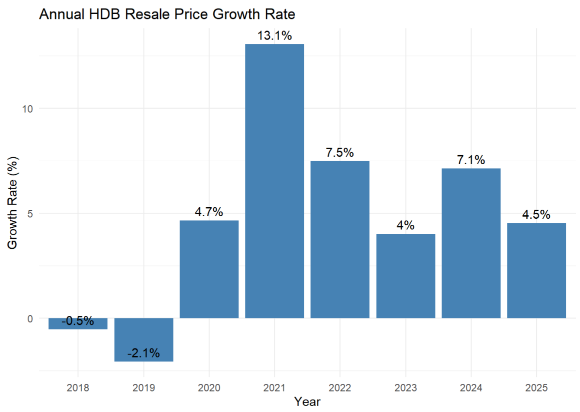


Figure 3: Annual Resale Price Growth Rate



Modelling (Target Variable: Resale Price)

Motivated to further investigate the increasing resale price, I attempted to build models for further exploration. Data Preprocessing Steps vary across models. For models that are sensitive to numeric values such as linear regression and neural network, I have scaled all numerical variables. While for tree-based models, I converted the date type variable **month** to numeric to ensure compatibility with the model structure. Besides, noticing the right-skewed distribution of the target variable, I applied log transformation for models like linear regression and natural splines.

Linear Regression (Baseline)

During implementation, I try to see how the dataset is away from model assumptions. I use variance inflation factor to assess possible multicollinearity and the potential collinearity between flat type and floor area is found. This is probably because those units with similar flat types would generally have similar ranges of floor areas. To decide whether to drop one of the terms or add an interaction term, an F test was conducted. It turns out that removing either variables would significantly worsen model performance. Furthermore, adding an interaction term significantly increased the model performance.

To further assess model performance, I build a studentized residual plot and notice the existence of many outliers, indicating a possible deficiency in the model. To investigate the possible non-linearity, I build residual plots across numeric variables, and discover potential nonlinearity for floor area (some curvature in the plot). Therefore, I consider an attempt of implementation of splines.

Natural Splines and Smoothing Splines

Both natural spline and smoothing spline are making attempts over the nonlinearity between resale price and floor area. While I conducted cross-validation to find a knot number that allows a good fit, the plot does not show any stable choice of degree of freedom. Therefore, to avoid underfitting and overfitting, I chose a degree of freedom of 7. I also attempted using GAM with smoothing splines. Yet, both natural splines and smoothing splines showed a minimal improvement from baseline. Therefore, the linear model remains a strong, interpretable, and efficient choice for predicting resale prices.

Applying ANOVA to all three models indeed showed me a statistically significant p-value for all 7 variables.

Regression Tree

To further address nonlinearity, I seek help from tree-based models. Compared to previous models that use all 7 variables, the single regression tree did not use the variable **storey_median**. However, this single regression tree performs worse than the baseline model both before and after pruning, indicating a high bias and insufficient model complexity. Given that our dataset contains a relatively small number of predictors, variance is not a major concern. Therefore, I focus on reducing bias using a boosting method.

Boosting and Extreme Gradient Boosting

Besides the boosting method taught in class, I attempted XGBoost considering its optimization algorithm. After tuning, the hyperparameter for boosting is (n.trees = 3000, interaction.depth = 4) and for XGBoost is (n.trees = 4394, max.depth = 6). Boosting provided variable importance measures, revealing **floor_area_sqm** as the most influential contributor.

Multilayer Neural Network (NN)

NN is usually considered risky (overfitting) for implementation for this kind of dataset, yet, I implemented it to retain categorical information for the variable **town** using an embedded layer. The network consists of two hidden layers of RELU with dropout. However, the model performs moderately.

Evaluation

Table 1: Overall Performance

Metrics	LR	NS	GAM SS	R Tree	Pru RT	Boosting	XGBoost	NN
RMSE	74762.49	74469.58	74585.93	102351.8	98982.57	50477.69	**49340.18**	94183.68
MAPE	10.53%	10.50%	10.51%	14.99%	14.49%	6.70%	**6.40%**	15.19%
R²	0.8262	0.8276	0.8271	0.6725	0.6937	0.9203	**0.9239**	0.7265

RMSE, MAPE, and R^2 values are based on the test set. I chose RMSE for its penalty on large errors, and MAPE for its easy interpretation. I did not use adjusted R^2 considering the value p is not really determined in complex models. While the XGBoost that optimizes the loss function with gradient descent performed the best, the baseline model unexpectedly achieved a relatively nice performance.

Table 2: Performance by flat type

Model	3 ROOM	4 ROOM	5 ROOM
lm.fit	61807.69	70665.52	90131.51
ns.fit	51074.13	70452.61	90054.73
gam.fit	51196.26	70440.81	90070.36
rg.fit	68840.87	99032.84	115976.88
boost.fit	32248.83	49798.10	60914.93
xgb.fit	**30190.90**	**47420.09**	**60214.17**
nn.fit	77543.00	88779.50	109721.21

To investigate which flat type was predicted most accurately, I evaluated model performance across the three categories with most samples: “3 ROOM,” “4 ROOM,” and “5 ROOM.” Among these, “3 ROOM” flats showed the lowest RMSE, likely due to its large sample size that reduces variance.

Conclusion and Reflection

I have observed that resale prices have consistently increased since 2020 and identified key variables that contribute to price prediction. While XGBoost presented strong predictive performance, the model can be further improved by incorporating additional features, such as the number of nearby MRT stations or schools; private property can also be considered. In the future, I aim to develop a price direction classification model, which may offer more practical value in real-world decision-making.

Appendix

Reference

Resale flat prices based on registration date from Jan-2017 onwards | HDB. (n.d.). data.gov.sg. https://data.gov.sg/datasets/d_8b84c4ee58e3cfc0ece0d773c8ca6abc/view

Phang, S., & Wong, W. (1997b). Government policies and private housing prices in Singapore. *Urban Studies*, *34*(11), 1819–1829. https://doi.org/10.1080/0042098975268

Wei, C. X. (2025, February 11). HDB resale prices set to continue growth streak in 2025: report. *The Business Times*. https://www.businesstimes.com.sg/property/hdb-resale-prices-set-continue-growth-streak-2025-report

Column Description

Original Dataset

Column Name	Data Type	Description	Example
month	object	Month of transaction (format: YYYY-MM)	2017-01
town	object	Town where the flat is located	ANG MO KIO
flat_type	object	Type of the flat	2 ROOM
block	object	Block number	406
street_name	object	Street name	ANG MO KIO AVE 10
storey_range	object	Range of storey the flat is located	10 TO 12

floor_area_sqm	float	Floor area in square meters	45.0
flat_model	object	Model type of the flat	Improved
lease_commence_date	int	Year when the lease started	1979
remaining_lease	object	Remaining lease duration	61 years 06 months
resale_price	float	Resale price in SGD	232000.0

Dataset after feature selection and engineering

Column Name	**Data Type**	**Description**	**Example**
month	object	Month of transaction (format: YYYY-MM)	2017-01
flat_type	object	Type of the flat	2 ROOM
storey_median	object	Median of the range of the storey	11
floor_area_sqm	float	Floor area in square meters	45.0
flat_model_group	object	Model type group of the flat	Basic
remaining_lease_years	object	Remaining lease duration in numeric representation	61.5 years
resale_price	float	Resale price in SGD	232000.0
region	object	Region of the flat	Central